SOME APPLICATIONS OF GENERALIZED INVERSE TO PATTERN RECOGNITION

A THESIS

Presented to

The Faculty of the Division of Graduate

Studies and Research

By

M. Adnan Al-Alaoui

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in the School of Electrical Engineering

Georgia Institute of Technology

November 1974

SOME APPLICATIONS OF GENERALIZED INVERSE

TO PATTERN RECOGNITION

Approved:

Roger P. Webb, Chairman

Aubrey M. Bush

Barry M. Leiner

James W. Walker

Edward M. Kamen

Date approved by Chairman 11/21/74

Dedicated to

My Parents

Mohamed Said Al-Alaoui, my father, *in Memoriam*

and

Souraya Hawasly Al-Alaoui, my mother

ACKNOWLEDGMENTS

I am grateful to my advisor, Professor Roger P. Webb, for his guidance and encouragement and to the members of the reading committee, Professors Aubrey M. Bush, Edward M. Kamen, Barry M. Leiner, M. Zuhair Nashed, and James W. Walker, for their constructive comments on the work. During the course of this work, I was also fortunate to have access to the counsels of Professor William G. Wee of the University of Cincinnati and Professors James E. Brown, Chee-Yee Chong, Atif S. Debs, William J. Kammerer, Michael D. Kelly, Mohamed F. Moad, and Monroe E. Womble of the Georgia Institute of Technology.

I am also indebted to Raman K. Sahgal for his help in Computer Programming. The manuscript was typed swiftly and accurately by Mrs. Betty C. Yarborough.

TABLE OF CONTENTS

TABLE OF CONTENTS (Concluded)

LIST OF TABLES

LIST OF ILLUSTRATIONS

SUMMARY

The purpose of this research is to apply the concepts and techniques of the generalized inverse to mean-square-error (MSE) problems in pattern recognition. The aim is to keep the attractive features of the MSE approach and try to combat its deficiencies, utilizing the attractive formulation of the generalized inverse and contributing new results. The scope of this dissertation is restricted to the more realistic case in pattern recognition where the underlying probability densities of the different classes are unknown. The results are particularly suitable to the equally realistic case of nonseparable classes.

The contributions of this research include:

1. Introducing a new weighted MSE procedure for pattern classification and motivating the approach statistically.

2. Three theorems on redundancy and the least-square generalized inverse solution for an inconsistent set of equations are presented and proved.

3. Introducing a new algorithm for pattern classification together with a convergence proof for the linearly separable case.

4. Introducing an adaptive constrained MSE procedure for pattern classification.

5. Suggesting a problem-oriented clustering technique.

6. Pointing out the relation between the generalized inverse, Fourier and Karhumen-Loéve expansions.

NOTATION

Unless otherwise defined, the following notations and typographical conventions are maintained:

†	- The Generalized Inverse
T	- Transpose
*	- Adjoint, Complex Conjugate Transpose
Upper Case Latin	- Matrices, Sets, Operators
Lower Case Latin	- Vectors
Lower Case Italics	- Scalars
Lower Case Greek	- Scalars.

Bibliographic references are enclosed in square brackets.

CHAPTER I

INTRODUCTION AND MATHEMATICAL FORMULATION

The purpose of this research is to apply the concepts and techniques of the generalized inverse to mean-square-error (MSE) problems in pattern recognition. MSE techniques are used extensively in pattern recognition for classification, feature extraction, and clustering. The generalized inverse approach to MSE yields a solution of minimum norm even when difficulties are encountered because of singularities. The scope of this dissertation is restricted to the more realistic case in pattern recognition where the underlying probability densities of the different classes are unknown.

In pattern classification the MSE criterion is statistically optimal under the Gaussian equal-covariance-matrix assumptions. The resulting pattern classifier is either linear, or its generalization a Φ machine [41]. The main attraction for using the MSE criterion is that we get a solution even when the classes are not linearly separable. The fact that the generalized inverse provides a closed form MSE solution was noted by Ho and Kashyap [25]. Wee [68], [69] applied the generalized inverse approach to multiclass pattern classification. Patterson and Womack [45] showed that the MSE solution gave a minimum-squared-error approximation to the Bayes discriminant and Wee [68] carried out the proof for the multiclass case. Unfortunately, this minimum-squared-error approximation to Bayes discriminant is weighted by the probability

of samples and thus emphasis is placed on points where the probability of the samples is large, rather than on points near the decision surface.

Several modifications of the MSE criterion are suggested in the literature of pattern recognition [Fukunaga (1972), p. 107] [17]. These modifications result in nonlinear functions and the explicit solutions which minimize these criteria are hard to obtain. Koford and Groner [30] have shown that the MSE classifier with the equal numbers of sample patterns for each class is equivalent to the linear optimal classifier if the patterns are Gaussian with equal-covariance matrices. They also showed that they get a nearly optimal classifier under the Gaussian equal-covariance assumptions if sample patterns for each class are not equal in their numbers, by weighting the MSE criterion by the inverse of the number of samples of each class. Wee [68] suggested weighting the samples of each class differently, but using the same weighting for samples in the same class. He did not suggest how to choose these weights and did not carry out any computations. While it is widely recognized that weighting will improve the MSE performance, in the absence of any probabilistic information the problem becomes how to choose the weights. Constrained MSE is used mainly to avoid a trivial solution [55]. For a survey of MSE in pattern recognition, see the paper by Yau and Garnett [73] and the recent book by Duda and Hart [13].

So far, the applications of the generalized inverse to pattern recognition have consisted mainly of the formulation of the problem in the generalized inverse setting [60], [68], [69], [70] with the resulting solution being the MSE solution of minimum norm. A notable exception

is the Ho-Kashyap algorithm [25] that yields a separable solution in the two-class pattern classification problem, if the patterns are linearly separable and gives an indication of nonseparability if the samples are not linearly separable.

The aim of this dissertation is to keep the attractive features of the MSE approach and to try to combat its deficiencies utilizing the attractive formulation of the generalized inverse, drawing on the vast literature of the generalized inverse, and contributing new results. In particular, the more difficult problem of nonseparable classes will be undertaken. The objective is to develop new techniques in pattern recognition utilizing the generalized inverse to the traditional purposes of pattern recognition in the nonparametric case. The purpose is reducing the error on the design set and getting better classification with the same number or fewer features than we get by the MSE solution. This will lead us into utilizing the more powerful concepts of generalized inverse such as the constrained least-square generalized inverse and the weighted least-square generalized inverse, and always giving us the corresponding unique solution of minimum norm. The current status in pattern classification reveals the existence of many algorithms for the separable case differing mainly in the rate of their convergence. The MSE criterion is popular because it yields a solution for both the separable and nonseparable case, and for the nonseparable case compares favorably with other algorithms [59].

Problem Formulation

The mathematical formulation of the pattern recognition problem

in this dissertation is based on the generalized inverse approach of Wee. The following mathematical formulation is used to describe the pattern recognition problem (Fig. 1).

1) Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ denotes the set of K pattern classes.

2) A set S denotes the training set consisting of N training samples from K classes with $\sum_{i=1}^{K} N_i = N$ where N_i designates the number of training samples from class i.

These training samples are denoted by the column vectors $y_j^{(i)}$ where i indexes the particular class and j indicates the j^{th} prototype from class i. Thus

$$y_j^{(i)T} = (y_{1j}^{(i)}, \ldots, y_{rj}^{(i)}, \ldots y_{Rj}^{(i)})$$

r ranges from 1 to R; j ranges from 1 to N_i; T denotes transpose.

3) Let Φ designate the feature extraction transformation that maps the primitive R-dimensional space to a lower P-dimensional space. The transformation can be written as a P-dimensional function acting on the primitive R-dimensional observations Y.

$$\Phi^T(\cdot) = [\varphi_1(\cdot), \ldots, \varphi_P(\cdot)]$$

Define the P-dimensional processed patterns

$$x_j^{(i)} = \Phi(y_j^{(i)}); \quad j = 1, 2, \ldots, N_i; \quad i = 1, 2, \ldots, K.$$

such that

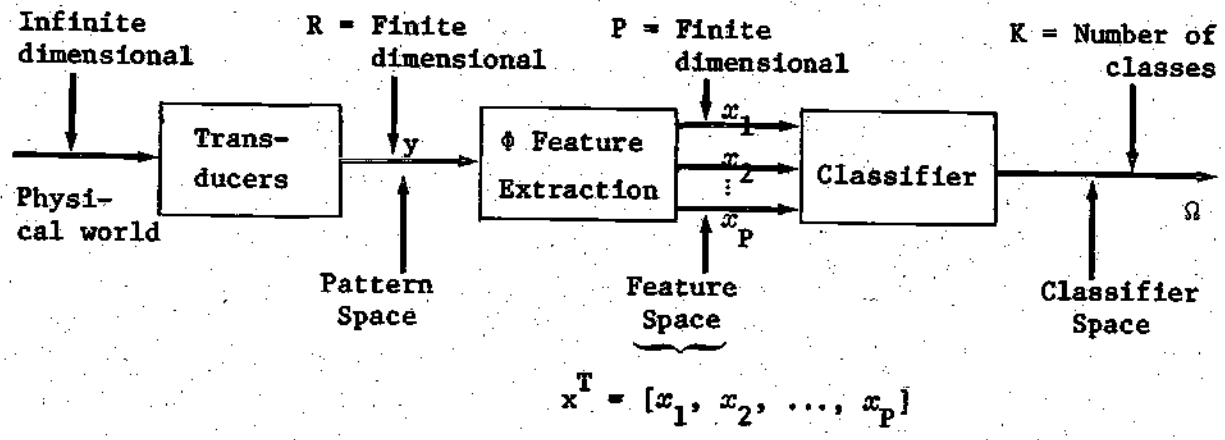


Figure 1. Overview of the Pattern Recognition Problem.

$$x_j^{(i)T} = [x_{1j}^{(i)}, \ldots, x_{Pj}^{(i)}] \qquad x_j^{(i)} \in \omega_i$$

where

$$x_{1j}^{(i)} = \varphi_1 (y_j^{(i)})$$

$$\vdots$$

$$x_{P-1,j}^{(i)} = \varphi_{P-1} (y_j^{(i)})$$

$$x_{Pj}^{(i)} = 1 \; \forall j \text{ and } \forall i$$

After the transformation of the training samples from y to x, the set S may be shown in matrix form as

$$A = \begin{bmatrix} A^{(1)} \\ A^{(2)} \\ \vdots \\ A^{(K)} \end{bmatrix} \qquad \text{where } A^{(i)} = \begin{bmatrix} x_1^{(i)T} \\ x_2^{(i)T} \\ \vdots \\ x_{N_i}^{(i)T} \end{bmatrix}$$

Thus A is an N x P matrix.

The Matrix A (N x P)

$$A = \begin{bmatrix} x_1^{(1)T} \\ \vdots \\ x_{N_1}^{(1)T} \\ x_1^{(2)T} \\ \vdots \\ x_{N_2}^{(2)T} \\ \vdots \\ x_1^{(K)T} \\ \vdots \\ x_{N_K}^{(K)T} \end{bmatrix} = \begin{bmatrix} x_{1,1}^{(1)} & \cdots & x_{P-1,1}^{(1)} & 1 \\ \vdots & & & \\ x_{1,N_1}^{(1)} & \cdots & x_{P-1,N_1}^{(1)} & 1 \\ x_{1,1}^{(2)} & \cdots & x_{P-1,N_1}^{(2)} & 1 \\ \vdots & & & \\ x_{1,N_2}^{(2)} & \cdots & x_{P-1,N_2}^{(2)} & 1 \\ \vdots & & & \\ x_{1,1}^{(K)} & \cdots & x_{P-1,1}^{(K)} & 1 \\ \vdots & & & \\ x_{1,N_K}^{(K)} & \cdots & x_{P-1,N_K}^{(K)} & 1 \end{bmatrix}$$

4) Let $\gamma(j/i)$ denote the cost incurred in classifying a pattern belonging to ω_i as ω_j.

$$c_i^T = [\gamma(1/i)\ \gamma(2/i)\ \ldots\ \gamma(K/i)]$$

Thus c_i is a K-dimensional vector.

A matrix B is defined as a set of cost vectors for each sample in S.

$$B = \begin{bmatrix} B_1 \\ B_2 \\ \cdot \\ \cdot \\ \cdot \\ B_K \end{bmatrix} \qquad \text{where } B_i = \begin{bmatrix} c_i^T \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ c_i^T \end{bmatrix}$$

B_i is an N_i x K matrix.

B is an N x K matrix where

$$N = \sum_{i=1}^{K} N_i \quad .$$

The Matrix B (N x K)

$$B = \begin{matrix} N_1 \text{ rows} \\ \\ \\ \\ \\ N_2 \text{ rows} \\ \\ \\ \\ \\ N_K \text{ rows} \\ \\ \\ \\ \end{matrix} \begin{bmatrix} \gamma(1/1) & \gamma(2/1) & \cdots & \gamma(K/1) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \gamma(1/1) & \gamma(2/1) & \cdots & \gamma(K/1) \\ \gamma(1/2) & \gamma(2/2) & \cdots & \gamma(K/2) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \gamma(1/2) & \gamma(2/K) & \cdots & \gamma(K/2) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \gamma(1/K) & \gamma(2/K) & \cdots & \gamma(K/K) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \gamma(1/K) & \gamma(2/K) & \cdots & \gamma(K/K) \end{bmatrix}$$

With equal cost of misrecognition, we have

$$\gamma(j/i) = \begin{cases} 0 & \text{if } i = j \\ \gamma > 0 & \text{if } i \neq j \end{cases}$$

which is the cost function that will be used in this thesis.

5) Let $D = d_i(x)$: $d_i(x) = x^T w^{(i)} = \omega_1^{(i)} x_1 + \omega_2^{(i)} x_2 + \ldots + \omega_{P-1}^{(i)} x_{P-1} + \omega_P^{(i)}$; $i = 1, 2, \ldots K$ denote the set of discriminant functions where

$$w^{(i)} = \begin{bmatrix} \omega_1^{(i)} \\ \vdots \\ \omega_P^{(i)} \end{bmatrix}$$

The pattern x is classified as belonging to ω_i if:

$$d_i(x) < d_j(x) \quad \text{for all} \quad j \neq i \quad .$$

The object of the pattern classification problem is to choose the discriminant function which will classify new samples of unknown category in such a way as to minimize the expected loss.

Note that the scalar term $\omega_P^{(i)}$ is added to the discriminant function for coordinate translation purposes. The matrix W is defined as:

$$W = \left[w^{(1)} \; w^{(2)} \; w^{(3)} \; \ldots \; w^{(K)} \right]$$

$$W = \begin{bmatrix} w_1^{(1)} & \cdots & w_1^{(K)} \\ w_2^{(1)} & & w_2^{(K)} \\ \vdots & & \\ w_P^{(1)} & & w_P^{(K)} \end{bmatrix}$$

W is a P x K matrix.

The central problem in pattern classification using linear discriminant functions is to determine the weights utilizing the training set of labeled samples. Suppose we have a set of N training samples, $y_1, \ldots, y_N$, some labeled ω_1 and some labeled ω_2. After the feature extraction stage we will have a set of N training samples, $x_1, \ldots, x_N$, some labeled ω_1 and some labeled ω_2. We want to use these samples to determine the weights in a linear discriminant function $d(x) = w^T x$. A sample x_j is classified correctly if $w^T x_j > 0$ and $x_j \in \omega_1$, or if $w^T x_j < 0$ and $x_j \in \omega_2$. In the latter case, we observe that x_j is classified correctly if $w^T(-x_j) > 0$. This suggests a normalization that simplifies the treatment of the two-category case, the replacement of all samples labeled ω_2, by their negatives. With this normalization we can forget the label and look for a weight vector w such that $w^T x_j > 0$ for all of the samples. Such a weight vector is called a solution vector. It should parenthetically be mentioned that a margin or a buffer zone can be defined to insure weight vector solutions which do not lie close to any prototype points in the pattern space. Thus,

$w^T x > b$, $b > 0$ would provide such a zone. Note that the normalization simplifying the two-category case makes many two-category techniques to find W inapplicable to the multiclass case.

For two classes the problem of finding d(x) that classifies all the given patterns correctly is equivalent to the problem of finding a solution to the vector inequality

$$Aw > 0$$

$$A = \begin{bmatrix} A^{(1)} \\ -A^{(2)} \end{bmatrix}$$

where $A^{(i)}$ is as defined previously.

A common procedure for solving linear inequalities is to transform the problem into an optimization problem, the solution of which also guarantees a solution for the inequality. The problem of minimizing a criterion function to determine the weight solution vector W can often be solved by gradient descent procedures. The chief concern here will be whether an iterative scheme does converge to the minimum of the criterion functions, and the rate of convergence [13].

Historically all the work on linear discriminant functions begins with the paper by R. A. Fisher (1936) [14]. The following table (Table 1) gives a summary of different criterion functions and descent procedures for obtaining a two-class linear discriminant function, (Duda and Hart, 1973) [13].

The Generalization Question

One of the basic problems of the algorithms, in the absence of

Table 1. Summary of Descent Procedures for Obtaining Linear Discriminant Functions [13].

Name	Criterion Function	Descent Algorithm	Conditions	Remarks
Fixed Increment	$J_p = \sum_{w^T x \leq 0} (-w^T x)$	$w_{k+1} = w_k + x_k;\ (w_k^T x_k \leq 0)$	———	Finite convergence if linearly separable to solution with $w^T x > 0$; w_k always bounded.
Variable Increment	$J_p' = \sum_{w^T x \leq b} -(w^T x - b)$	$w_{k+1} = w_k + \rho_k x_k;\ (w_k^T x_k < b)$	$\rho_k \geq 0, \Sigma\rho_k \to \infty, \frac{\Sigma\rho_k^2}{(\Sigma\rho_k)^2} \to 0$	Convergence if linearly separable to solution with $w^T x > b$. Finite convergence if $0 < \alpha \leq \rho_k \leq \beta < \infty$
Relaxation	$J_r = \frac{1}{2} \sum_{w^T x \leq b} \frac{(w^T x - b)^2}{\|x\|^2}$	$w_{k+1} = w_k + \rho \frac{b - w_k^T x_k}{\|x_k\|^2} x_k$ $(w_k^T x_k \leq b)$	$0 < \rho < 2$	Convergence if linearly separable to solution with $w^T x \geq b$. If $b > 0$, finite convergence to solution with $w^T x > 0$
Widrow-Hoff	$\frac{1}{2} J_s = \frac{1}{2} \sum (w^T x_i - b_i)^2$	$w_{k+1} = w_k + \rho_k (b_k - w_k^T x_k) x_k$	$\rho_k > 0,\ \rho_k \to 0$	Tends toward solution minimizing J_s.
Stochastic Approximation	$J_m = \epsilon[(w^T x - z)^2]$	$w_{k+1} = w_k + \rho_k (z_k - w_k^T x_k) x_k$	$\Sigma\rho_k \to \infty\ \ \Sigma\rho_k^2 \to L < \infty$	Involves an infinite number of randomly drawn samples; converges in mean square to a solution minimizing J_m; also provides a MSE approximation to Bayes discriminant.
		$w_{k+1} = w_k + R_k (z_k - w_k^T x_k) x_k$	$R_{k+1}^{-1} = R_k^{-1} + x_k x_k^T$	

Table 1 (Concluded)

Name	Criterion Function	Descent Algorithm	Conditions	Remarks
Pseudo-Inverse	$J_s = \|\|Aw - b\|\|^2$	$w = A^{\dagger}b$	———	Classical MSE solution: special choices for b yield Fisher's linear discriminant and MSE approximation to Bayes discriminant.
Ho-Kashyap	$J_s = \|\|Aw - b\|\|^2$	$b_{k+1} = b_k + \rho(e_k + \|e_k\|)$ $e_k = Aw_k - b_k$ $w_k = A^{\dagger}b_k$	$0 < \rho < 1$, $b_o > 0$	w_k is MSE solution for each b_k; finite convergence if linearly separable; if $e_k \leq 0$ but $e_k \neq 0$, the samples are nonsparable.
		$b_{k+1} = b_k + (e_k + \|e_k\|)$ $w_{k+1} = w_k + \rho_k R A^T \|e\|$	$\rho_k = \dfrac{\|e_k\|^T ARA^T \|e_k\|}{\|e_k\|^T ARA^T ARA^T \|e_k\|}$ is optimum R symmetric, positive definite; $b_o > 0$	Finite convergence if linearly separable; if $A^T\|e_k\| = 0$ but $e_k \neq 0$ the samples are nonseparable.
Linear Programming	$t = \max [-(w^T x_i - b_i)]$ $w^T x_i \leq b_i$	Simplex algorithm	$w^T x_i + t \geq b_i$, $t \geq 0$	Finite convergence in both separable and nonseparable cases; useful solution only if separable.
	$J_p = \sum_{i=1}^{n} t_i$ $= \sum_{w^T x_i \leq b_i} -(w^T x_i - b_i)$	Simplex algorithm	$w^T x_i + t_i \geq b_i$, $t_i \geq 0$	Finite convergence to solution minimizing perceptron criterion function whether separable or not.

any probabilistic information, is the question of generalization. The only result along this line seems to be the recent important result by Foley [15]. Foley shows that in general the number of samples N must be equal to or larger than four times the number of features for the algorithms of this case to yield meaningful results. For the mean-square-error (MSE) criterion, Patterson and Womack [45] showed that the MSE solution gave a minimum-squared-error approximation to the Baye's discriminant. Unfortunately, this minimum-squared-error approximation to Bayes discriminant is weighted by the probability of the samples and thus emphasis is placed on points where the probability of the samples is large rather than on points near the decision surface.

Multiclass Problem

The multiclass problem may be of three forms as follows [23]:

1) Each class may be separable from all the rest by a single decision surface. Then, we may take the decision according to

$$d_i(x) > 0 \qquad \text{if } x \in \omega_i$$

$$< 0 \qquad \text{otherwise}$$

this reduces the multiclass problem to K-1 two-class problems.

2) Each class may be separable from each other class. Here we have $\frac{K(K-1)}{2}$ two-class problems and as many decision functions such that

$$d_{ij}(x) > 0 \qquad \text{if } x \in \omega_i$$

$$< 0 \qquad \text{if } x \in \omega_j$$

given that x belongs to ω_i or ω_j.

3) There exists K $d_i(x)$ such that x belongs to ω_i only if

$$d_i(x) > d_j(x) \qquad \text{for all } j \neq i \ .$$

Note that this is equivalent to case 2) as we may define

$$d_{ij}(x) = d_i(x) - d_j(x) \ .$$

The MSE criterion for multiclass problem is of type 3).

If each class is separable from all the rest by a single decision surface, then the generalization of the procedures from a two-class to multiclass problems is straightforward. However, if each class is not separable from all the rest by a single decision surface, then some of the two-category procedures cannot be extended to the multiclass case.

Nonseparable Behavior

In most pattern classification applications one cannot assume that the samples are linearly separable. In particular, when the patterns are not separable, one still wants to obtain a weight vector that classifies as many samples correctly as possible. An objective function whose minimization would minimize the number of classification errors is

$$J(w) = \frac{c_1}{N_1} \sum_{j=1}^{N_1} J_j^1 \left(x_j^{(1)^T} w\right) + \frac{c_2}{N_2} \sum_{j=1}^{N_2} J_j^2 \left(x_j^{(2)^T} w\right)$$

$$J_j^i\left(x_j^{(i)^T} w\right) = 1 \ , \quad x_j^{(1)^T} w \leq 0 \ , \quad x_j^{(2)^T} w \geq 0$$

$$= 0 \quad \text{otherwise}$$

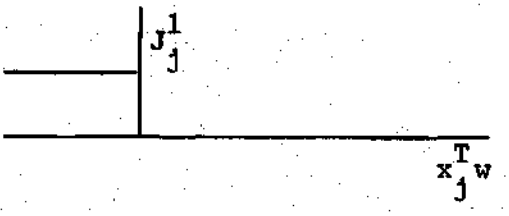


Figure 2. Pattern Error Function for Minimun Number of Classification Errors.

The mean of the resulting objective function averaged over all x is the probability of error assuming that c_1 and c_2 are set to the a priori probabilities of the categories. The pattern error function is shown in Fig. 2. The function is nonconvex and efforts to minimize it will have to contend with relative minima as well as discontinuties.

The exact nonseparable behavior of the different algorithms has been studied thoroughly in only a few special cases. It is known, for example, that the length of the weight vectors produced by the fixed increment rule is bounded. Emperical rules for terminating the correction procedure are often based on this tendency of the weight vector to fluctuate near some limiting value. The MSE solution yields a solution for both the linearly separable and nonseparable case. Although the MSE does not necessarily yield a separating solution in the linearly separable case, it is reasonable to hope that by minimizing the MSE criterion function we might obtain a useful discriminant function in both the separable and the nonseparable cases. Except for very special cases, it is impossible so far to obtain analytical results on the expected loss of the pattern classifiers based on the least-mean-square approach. This problem seems to be the most well-known unsolved

problem in this area [73].

Thesis Outline

Chapter II deals with weighted MSE in the generalized inverse setting. The contribution of this dissertation in Ch. II includes:

1) Introducing a new weighting scheme and motivating it statistically.

2) Three theorems on redundancy and the least-square generalized inverse solution for an inconsistent set of equations are presented and proved.

3) The method is carried out on the data base of Sebestyen and Edie [54] with very favorable results.

Chapter III presents the algorithm resulting from the method in Ch. II. The contributions in Ch. III include:

1) Presentation of a new algorithm for pattern classification together with a convergence proof for the linearly separable case.

2) A comparison of the new algorithm with the relaxation and the Ho-Kashyap algorithms.

3) The new algorithm is carried out on several linearly separable examples.

Chapter IV deals with constrained MSE. The contributions of this dissertation in Ch. IV include:

1) Introducing an adaptive constrained MSE procedure.

2) Applying the constrained procedure to Sebestyen and Edie's data with very favorable results.

Chapter V deals with the weighted means. In the chapter the samples in a class are represented by their means and the weighting

techniques of Ch. II is applied. The contributions in this chapter include:

1) Suggesting a problem-oriented clustering technique.

2) Applying the weighted-means technique to Sebestyen and Edie's data with favorable results.

Chapter VI deals with feature extraction. The contribution of this dissertation in Ch. VI is pointing out the relation between the generalized inverse, Fourier and Karhunen-Loéve expansions.

Chapter VII deals with some computational aspects. In particular it details Kishi's algorithm which is particularly suitable to the techniques of Ch. II and Ch. III. The contribution of this dissertation in Ch. VII is mainly editorial.

Two appendices are provided. Appendix A provides a quick reference to the properties of the generalized inverse. Appendix B presents Sebestyen and Edie's data.

CHAPTER II

APPLICATION OF WEIGHTED GENERALIZED INVERSE TO PATTERN CLASSIFICATION

In this chapter a new method of weighted MSE is presented. The method is motivated by Patterson and Womack result [45] that the MSE criterion provides a mean-square-error approximation to Bayes' discriminant weighted by the probability of the samples. To offset this weighting and emphasize the patterns that are away from the mode, the misclassified samples are repeated to increase their probability. This leads to three new theorems on redundancy and the MSE solutions which are presented in this chapter. The first theorem shows that repeating a row in a system of an inconsistent set of equations reduces the error for the repeated row. The second theorem shows that repeating a row in a system of an inconsistent set of equations results in a solution that could be expressed as a weighted MSE solution of the original system and hence the title of this chapter. The third theorem shows that repeating a row in an inconsistent set of equations is equivalent to changing b in the system of equations $Aw = b$ in the direction of the gradient of the norm $||Aw - b||^2$ with respect to b. Lemma 1 shows that repeating a row in an inconsistent set of equations, utilizing Theorem 3, is equivalent to increasing the cost of misrecognition for the repeated sample.

Smith [59] showed in a comparison of the MSE solution with the fixed-increment and relaxation methods on linearly nonseparable samples

that the MSE solution gave the least number of errors. The method presented in this chapter of iteratively repeating the misclassified samples gives considerably better results than the MSE solution as demonstrated by the example presented at the end of this chapter. The method also yields a new algorithm that converges to a separating solution in a finite number of steps if the patterns are linearly separable. The algorithm is presented in the next chapter.

In least-squares estimate [32], [53] we want to solve for W that minimizes $||AW - B||$. Assuming A is of full column rank the solution is:

$$\hat{W} = A^{\dagger}B = (A^{T}A)^{-1}A^{T}B$$

which is the solution of minimum norm among all possible solutions.

By weighted least-squares we mean minimizing

$$(AW - B)^{T} R^{-1}(AW - B) = ||AW - B||^{2}_{R^{-1}}$$

where T denotes transpose and R is a positive definite matrix, hence there exists a matrix Q such that $Q^{T}Q = R^{-1}$. To do this it is only necessary to consider a matrix equation of the form QAW = QB + QE instead of AW = B + E. Hence, the least squares solution for W is given by the so-called weighted generalized inverse solution [9]

$$\hat{W} = (QA)^{\dagger} QB \quad .$$

If A is of full column rank, this becomes [53]

$$\hat{W} = (A^{T}R^{-1}A)^{-1} A^{T}R^{-1}B \quad .$$

The point of departure of the weighting approach is the Gauss-Markov

theorem.

Theorem (Gauss-Markov) suppose $AW = B + E$ where

$$\varepsilon(E) = 0$$

$$\varepsilon(EE^T) = R$$

(where ε = Expected Value).

With R positive definite, the linear minimum-variance unbiased estimate of W is:

$$\hat{W} = (A^T R^{-1} A)^{-1} A^T R^{-1} B \quad .$$

A striking property of the result of Gauss-Markov theorem is that if $\varepsilon(EE^T) = I$, the linear, minimum-variance unbiased estimate is identical to the least-squares estimate.

In the absence of any probabilistic information, a suggested method for least-square-error refinement is to estimate R by the sample error covariance matrix and carry this repeatedly. If A is m x n matrix with $m > n$, then R would be m x m. Since we have to find R^{-1}, this would entail inverting an m x m matrix. This prospect is not very attractive in pattern recognition since m tends to be very large.

Statistical Justification of the Method

Patterson and Womack [45] showed that the MSE solution gave a minimum-square-error approximation to the Bayes discriminant and Wee [68] carried out the proof for the multiclass case. We present here the proof for the 2 classes since it is pertinent.

Asymptotic Approximation of MSE Solution to Bayes

Bayes Discriminant Function (2 classes)

$$d_o(x) = P(\omega_1|x) - P(\omega_2|x) \quad .$$

Assumption. The samples are drawn independently according to the probability law:

$$P(x) = P(x|\omega_1)P(\omega_1) + P(x|\omega_2)P(\omega_2) \quad .$$

The Criterion Function

$$J(w) = \sum_{x\epsilon\omega_1} (w^T x - 1)^2 + \sum_{x\epsilon\omega_2} (w^T x + 1)^2$$

$$J(w) = N\left[\frac{N_1}{N} \cdot \frac{1}{N_1} \sum_{x\epsilon\omega_1} (w^T x - 1)^2 + \frac{N_2}{N} \cdot \frac{1}{N_2} \sum_{x\epsilon\omega_2} (w^T x + 1)^2\right]$$

By the law of large numbers, as N approaches infinity, $\frac{1}{N} J(W)$ approaches:

$$\bar{J}(w) = P(\omega_1)\varepsilon_1[(w^T x - 1)^2] + P(\omega_2)\varepsilon_2[(w^T x + 1)^2]$$

with probability one, where

$$\varepsilon_1[(w^T x - 1)^2] = \int (w^T x - 1)^2 \, P(x|\omega_1) \, dx$$

$$\varepsilon_2[(w^T x + 1)^2] = \int (w^T x + 1)^2 \, P(x|\omega_2) \, dx$$

Rewrite Bayes discriminant as:

$$d_o(x) = \frac{P(x,\omega_1) - P(x,\omega_2)}{P(x)} \quad .$$

Hence:

$$\tilde{J}(w) = \int (w^T x - 1)\ P(x,\omega_1)dx + \int (w^T x + 1)^2\ P(x,\ \omega_2)\ dx$$

$$\tilde{J}(w) = \int (w^T x)^2\ P(x)dx - 2 \int w^T x\ d_o(x)\ P(x)dx + 1$$

$$\boxed{\tilde{J}(w) = \int [w^T x - d_o(x)]^2\ P(x)dx + [1 - \int d_o^2(x)\ P(x)\ dx]}$$

The second term is independent of w. Hence, the w that minimizes J also minimizes ε^2, the mean-squared-error between $w^T x$ and $d_o(x)$.

$$\varepsilon^2 = \int [w^T x - d_o(x)]^2\ P(x)\ dx$$

Thus, the mean-square-error criterion places emphasis on points where P(x) is large, rather than on points near the decision surface $d_o(x) = 0$.

By repeating the misclassified samples we are in effect increasing their probabilities and thus placing the emphasis where it should be placed, on points near the decision surface.

This suggests an iterative procedure which starting with the MSE solution repeats the misclassified samples then tests the resulting solution on all the samples and repeats the resulting misclassified samples and so on. The procedure could be terminated either after the error was reduced to a certain value or by specifying the number of iterations to be carried out and keeping the one that results with the least number of errors.

In the following pages we present three theorems on the effect of repeating a sample. The first shows that the error for the repeated sample is reduced. The second shows that repeating a sample is equivalent

to weighting. The third shows that repeating a sample corresponds in the problem of minimizing $||Aw - b||$ to the following alternate steps: 1) For a fixed b a vector is determined such that it constitutes a least square fit. 2) For a fixed w, the component of b that corresponds to the repeated sample is changed in the direction of the gradient of the norm with respect to b.

Theorem 1

Let $Aw = b$ be an inconsistent set of equations where $A = \begin{pmatrix} a_1^* \\ \vdots \\ a_K^* \end{pmatrix}$ where each a_i^* corresponds to a row of A. Then the minimum norm solution is $\hat{w} = A^\dagger b$. Let $\tilde{A} = \begin{pmatrix} A \\ a_i^* \end{pmatrix}$, $\tilde{b} = \begin{pmatrix} b \\ b_i \end{pmatrix}$ where a_i^* and b_i correspond to repeating a row in the original A and b. Then the minimum norm solution for $\tilde{A}\tilde{w} = \tilde{b}$ is:

$$\hat{\tilde{w}} = \tilde{A}^\dagger \tilde{b}$$

and

$$(a_i^* \hat{\tilde{w}} - b_i) < (a_i^* \hat{w} - b_i)$$

Proof

The proof utilizes Kishi's algorithm [29] which is presented below. Let

$$\underset{Kx1}{b_K} = \underset{Kxn}{A_K} \underset{nx1}{w_K}$$

then the MSE solution of minimum norm is

$$\underset{nx1}{\hat{w}_K} = \underset{nxK}{A_K^\dagger} \underset{Kx1}{b_K} .$$

Define c_{K+1}^*, A_{K+1} and b_K as follows

$$A_{K+1} = \begin{pmatrix} A_K \\ a_i^* \end{pmatrix}$$

$$c_{K+1}^* = a_i^* - a_i^* A_K^\dagger A_K$$

$$b_{K+1} = \begin{pmatrix} b_K \\ b_i \end{pmatrix}$$

Case 1

$c_{K+1} \neq 0$:

$$h_{K+1} = c_{K+1}(c_{K+1}^* c_{K+1})^{-1}$$

Case 2

$c_K = 0$: $\begin{cases} \text{i) } A_K \text{ is of full column rank, and/or} \\ \text{ii) } a_i^* \text{ is in the row space of } A_K \end{cases}$

$$h_{K+1} = (1 + a_i^* A_K^\dagger A_K^{\dagger *} a_i)^{-1} A_K^\dagger A_K^{\dagger *} a_i \quad .$$

For both cases we have the following relations:

$$\boxed{\hat{w}_{K+1} = \hat{w}_K - h_{K+1} a_i^* \hat{w}_K + h_{K+1} b_i}$$

$$A_{K+1}^\dagger A_{K+1} = A_K^\dagger A_K + h_{K+1} c_{K+1}^*$$

$$A_{K+1}^\dagger A_{K+1}^{\dagger *} = (h_{K+1} a_i^* - I) A_K^\dagger A_K^{\dagger *} (I - (h_{K+1} a_i^*)^*) + h_{K+1} h_{K+1}^*$$

$$a_i^* \hat{w}_{K+1} - b_i = a_i^* \hat{w}_K - a_i^* h_{K+1} a_i^* \hat{w}_K + a_i^* h_{K+1} b_i - b_i$$

$$= (1 - a_i^* h_{K+1}) a_i^* \hat{w}_K - (1 - a_i^* h_{K+1}) b_i$$

hence, $$\boxed{a_i^* \hat{w}_{K+1} - b_i = (1 - a_i^* h_{K+1})\ (a_i^* \hat{w}_K - b_i)}$$

We could confine our investigation to the factor $1 - a_i^* h_{K+1}$. The case of interest to us is case 2.

Case 2

$$c_{K+1} = 0 \Longrightarrow a_i^* = a_i^* A_K^\dagger A_K$$

$$h_{K+1} = (1 + a_i^* A_K^\dagger A_K^{\dagger *} a_i)^{-1} A_K^\dagger A_K^{\dagger *} a_i$$

$$a_i^* h_{K+1} = \frac{a_i^* A_K^\dagger A_K^{\dagger *} a_i}{1 + a_i^* A_K^\dagger A_K^{\dagger *} a_i}$$

$$1 - a_i^* h_{K+1} = \frac{1}{1 + a_i^* A_K A_K^* a_i}$$

$$= \frac{1}{1 + (a_i^* A_K^\dagger)(a_i^* A_K^\dagger)^*}$$

$$\leq 1$$

with equality if $a_i^* = 0$ or a_i orthogonal to the column space of $A_K^\dagger$, i.e., orthogonal to the column space of A_K^T, i.e., orthogonal to the transposed row space of A_K, but by assumption a_i^* is in the row space of A_K · Q.E.D.

Example

$$A = \begin{pmatrix} 1 & 1 \\ & \\ 1 & 1 \end{pmatrix}, \qquad w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$||Aw - b||^2 = (w_1 + w_2 - 1)^2 + (w_1 + w_2 - 2)^2$$

$$\therefore \quad \frac{\partial}{\partial w_1} ||Aw - b||^2 = \frac{\partial}{\partial w_2} ||Aw - b||^2 = 2w_1 + 2w_2 - 3 = 0$$

$$\Longrightarrow \boxed{w_1 + w_2 = \frac{3}{2}} \quad \text{for a minimum mean square error.}$$

The minimum of the sum of the squares of residuals is:

$$\min_{w} ||Aw - b||^2 = (\tfrac{3}{2} - 1)^2 + (\tfrac{3}{2} - 2) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \quad .$$

Suppose we want to reduce the error in the first factor, we propose that introducing redundancy by repeating that factor twice in the original system, the MSE procedure will try to adjust the solution to reduce the error of the first factor.

Let

$$B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}, \qquad v = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

$$Bw - v = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} w_1 + w_2 - 1 \\ w_1 + w_2 - 2 \\ w_1 + w_2 - 1 \end{pmatrix}$$

$$||Bw - v||^2 = (w_1 + w_2 - 1)^2 + (w_1 + w_2 - 2)^2 + (w_1 + w_2 - 1)^2$$

$$\frac{\partial}{\partial w_1} ||Bw - v||^2 = \frac{\partial}{\partial w_2} ||Bw - v||^2 = 2(w_1 + w_2 - 1) + 2(w_1 + w_2 - 2)$$

$$+ 2(w_1 + w_2 - 1) = 0$$

$$\Longrightarrow 3w_1 + 3w_2 - 4 = 0 \qquad \boxed{w_1 + w_2 = \frac{4}{3}} \quad .$$

Substituting $w_1 + w_2 = \frac{4}{3}$ instead of $w_1 + w_2 = \frac{3}{2}$ in

$$||\bar{Aw} - b||^2 = (w_1 + w_2 - 1)^2 + (w_1 + w_2 - 2)^2$$

$$= (\tfrac{4}{3} - 1)^2 + (\tfrac{4}{3} - 2)^2$$

$$= (\tfrac{1}{3})^2 + (\tfrac{2}{3})^2$$

$$= \tfrac{1}{9} + \tfrac{4}{9} = \tfrac{5}{9} \quad .$$

The effect was to reduce the square of the error of the first factor from $\frac{1}{4}$ to $\frac{1}{9}$ and to increase the square of the error of the second factor from $\frac{1}{4}$ to $\frac{4}{9}$. Hence the effect is the same as introducing a weighting matrix.

Suppose we wanted to reduce the error in the second factor.

Let

$$B_1 = B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \quad , \quad w = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad , \quad v_1 = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}$$

$$\therefore \; ||B_1 w - v_1||^2 = (w_1 + w_2 - 1)^2 + 2(w_1 + w_2 - 2)^2$$

$$\frac{\partial}{\partial w_1}||B_1 w - v_1||^2 = \frac{\partial}{\partial w_2}||B_1 w - v_1||^2 = 2(w_1 + w_2 - 1)$$

$$+ 4(w_1 + w_2 - 2) = 0$$

$$\Longrightarrow 3w_1 + 3w_2 - 5 = 0 \qquad \therefore \quad \boxed{w_1 + w_2 = \frac{5}{3}} \ .$$

Substituting $w_1 + w_2 = \frac{5}{3}$ in $||Aw - b||^2$ we get:

$$||Aw - b||^2 = (w_1 + w_2 - 1)^2 + (w_1 - w_2)^2$$

$$= (\tfrac{5}{3} - 1)^2 + (\tfrac{5}{3} - 2)^2$$

$$= (\tfrac{2}{3})^2 + (\tfrac{1}{3})^2$$

$$= \tfrac{4}{9} + \tfrac{1}{9} = \tfrac{5}{9}$$

Compare with $\min ||Aw - b||^2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

Hence, the effect was to reduce the error of the second factor from $\frac{1}{4}$ to $\frac{1}{9}$ and increase the error in the first factor from $\frac{1}{4}$ to $\frac{4}{9}$. It is obvious that if we repeated both factors the same number of times that our solution will be the same as the original solution.

Theorem 2

Let $Aw = b$ be an inconsistent set of equations where

$$A = \begin{pmatrix} a_1^* \\ \vdots \\ a_K^* \end{pmatrix}$$

where each a_i corresponds to a row of A. Let $\tilde{A} = \begin{pmatrix} A \\ a_i^* \end{pmatrix}$, $\tilde{b} = \begin{pmatrix} b \\ b_i \end{pmatrix}$ where a_i^* and b_i correspond to repeating a row in the original A and b. Then

the minimum norm solution for $\tilde{A}w = \tilde{b}$ is $\hat{\tilde{w}} = \tilde{A}^{\dagger}\tilde{b}$ is equal to the solution of the system $||Aw - b||_R^2 = (Aw - b)^T R(Aw - b)$; i.e., a weighted MSE solution with $R = \begin{bmatrix} I & 0 \\ 0 & 2 \end{bmatrix}$ if we repeated the last row.

Proof

Let

$$A = \begin{pmatrix} \bar{A} \\ a_K^* \end{pmatrix} , \quad \tilde{A} = \begin{pmatrix} \bar{A} \\ a_K^* \\ a_K^* \end{pmatrix} , \quad \tilde{b} = \begin{pmatrix} \bar{b} \\ b_K \\ b_K \end{pmatrix}$$

$$||\tilde{A}w - \tilde{b}||^2 = \left[\begin{pmatrix} \bar{A} \\ a_K^* \\ a_K^* \end{pmatrix} w - \begin{pmatrix} \bar{b} \\ b_K \\ b_K \end{pmatrix}\right]^T \left[\begin{pmatrix} \bar{A} \\ a_K^* \\ a_K^* \end{pmatrix} w - \begin{pmatrix} \bar{b} \\ b_K \\ b_K \end{pmatrix}\right]$$

$$= \begin{pmatrix} \bar{A}w - \bar{b} \\ a_K^* w - b_K \\ a_K^* w - b_K \end{pmatrix}^T \begin{pmatrix} \bar{A}w - \bar{b} \\ a_K^* w - b_K \\ a_K^* w - b_K \end{pmatrix}$$

$$= [(\bar{A}w - \bar{b})^T \quad (a_K^* w - b_K)^T \quad (a_K^* w - b_K)^T] \begin{bmatrix} \bar{A}w - \bar{b} \\ a_K^* w - b_K \\ a_K^* w - b_K \end{bmatrix}$$

$$= (\bar{A}w - \bar{b})^T (\bar{A}w - \bar{b}) + 2(a_K^* w - b_K)^T (a_K^* w - b_K)$$

On the other hand we have:

$$(Aw - b)^T \begin{bmatrix} I & 0 \\ 0 & 2 \end{bmatrix} (Aw - b)$$

$$= \begin{pmatrix} \bar{A}w - \bar{b} \\ a_K^* w - b_K \end{pmatrix}^T \begin{bmatrix} I & 0 \\ 0 & 2 \end{bmatrix} \begin{pmatrix} \bar{A}w - \bar{b} \\ a_K^* w - b_K \end{pmatrix}$$

$$= [(\bar{A}w - \bar{b})^T \quad (a_K^* w - b_K)^T] \begin{bmatrix} I & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} Aw - \bar{b} \\ a_K^* w - b_K \end{bmatrix}$$

$$= [\bar{A}w - \bar{b})^T \quad 2(a_K^* w - b_K)^T] \begin{bmatrix} \bar{A}w - \bar{b} \\ a_K^* w - b_K \end{bmatrix}$$

$$= (\bar{A}w - \bar{b})^T \ (\bar{A}w - \bar{b}) + 2(a_K^* w - b_K)^T \ (a_K^* w - b_K)$$

$$\therefore \quad ||Aw - b||_R^2 = ||\bar{A}w - \bar{b}||^2$$

Q. E. D.

Theorem 3

Let $A_K w = b_K$ be a set of equations where

$$A_K = \begin{pmatrix} a_1^* \\ \vdots \\ a_K^* \end{pmatrix} ,$$

a_i^* corresponds to the i^{th} row of A_K and $b_K = \begin{pmatrix} b_1 \\ \vdots \\ b_K \end{pmatrix}$. The minimum norm solution $\hat{\tilde{w}} = \hat{A}_K^{\dagger} \tilde{b}_K$ resulting from repeating the i^{th} row is equal to the minimum norm soluiton resulting from changing the corresponding b_i in the direction of the gradient of $||A_K w - b_K||$ with respect to b_i.

Proof

Let a_i^* be the repeated sample, then the solution vector is

$$\hat{\tilde{w}} = \hat{w}_{K+1} = \hat{w}_K - h_{K+1} \ a_i^* \ \hat{w}_K + h_{K+1} \ b_i$$

where

$$\hat{w}_K = A_K^{\dagger} b_K$$

$$h_{K+1} = \frac{A_K^{\dagger} A_K^{\dagger *} a_1}{1 + a_1^* A_K^{\dagger} A_K^{\dagger *} a_i} = \rho_K A_K^{\dagger} A_K^{\dagger *} a_i$$

where

$$\rho_K = \frac{1}{1 + a_i^* A_K^{\dagger} A_K^{\dagger *} a_i} \leq 1 \quad .$$

From the theory of generalized inverse (Penrose) [46]

$$A_K^{\dagger} A_K^{\dagger *} = (A_K^* A_K)^{\dagger}$$

$$A_K^{\dagger} = (A_K^* A_K)^{\dagger} A_K^*$$

where

$$A_K = \begin{bmatrix} a_1^* \\ \vdots \\ a_K^* \end{bmatrix} \quad ; \quad A_K^* = [a_1 \ . \ . \ . \ a_K]$$

$$\therefore \quad A_K^{\dagger} = [(A_K^* A_K)^{\dagger} a_1 \ . \ . \ . \ (A_K^* A_K)^{\dagger} a_K]$$

$$\hat{w}_K = A_K^{\dagger} b_K = [(A_K^* A_K)^{\dagger} a_1 \ . \ . \ . \ (A_K^* A_K)^{\dagger} a_K] \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix}$$

$$\therefore \quad \hat{w}_K = (A_K^* A_K)^{\dagger} a_1 b_1 + \ . \ . \ . \ + (A_K^* A_K)^{\dagger} a_K b_K \qquad (1)$$

$$\hat{w}_{K+1} = \hat{w}_K - h_{K+1} [a_i^* \hat{w}_K - b_i]$$

$$\hat{w}_{K+1} = \hat{w}_K - \rho_K (A_K^* A_K)^\dagger a_i [a_i^* \hat{w}_K - b_i] \qquad (2)$$

Substituting the value of $\hat{w}_K$ from (1) in (2) we get:

$$\hat{w}_{K+1} = (A_K^* A_K)^\dagger a_1 b_1 + \ldots + (A_K^* A_K)^\dagger a_i b_i$$

$$+ \ldots + (A_K^* A_K)^\dagger a_K b_K - \rho_K (A_K^* A_K)^\dagger a_i [a_i^* \hat{w}_K - b_i]$$

$$\therefore \quad \hat{w}_{K+1} = (A_K^* A_K)^\dagger a_1 b_1 + \ldots + (A_K^* A_K)^\dagger a_i [b_i - \rho_K (a_i^* \hat{w}_K - b_i)]$$

$$+ \ldots + (A_K^* A_K)^\dagger a_K b_K$$

which is the solution resulting from substituting for b_i the quantity:

$$b_i - \rho_K (a_i^* \hat{w}_K - b_i) \quad .$$

It is well known that the gradients of $J = \frac{1}{2} ||Aw - b||^2$ with respect to w and b are:

$$\frac{\partial J}{\partial w} = A^T (Aw - b) \Longrightarrow \quad w = (A^T A)^\dagger A^T b \triangleq A^\dagger b$$

$$\frac{\partial J}{\partial b} = (b - Aw) \qquad \text{and} \qquad \frac{\partial J}{\partial b_i} = b_i - a_i^* w \quad .$$

Hence repeating a row that is in error consists of the following alternate steps:

1) For a fixed b, a vector w is determined such that it constitutes a least-square fit.

2) For a fixed w, allow the component of b that corresponds to the sample to be repeated to change in the direction of the gradient with respect to b.

Theorem 3 is utilized in Lemma 1, and in Ch. III in the comparison of the algorithm resulting from this procedure with the Ho-Kashyap algorithm.

Lemma 1: Repeating a misclassified sample is equivalent to increasing the cost of its misclassification with the class in which it is erroneously classified.

Proof:

For a two-class problem the formulation

$$\min_{w} ||A_1 w - b|| \qquad \text{where } A_1 = \begin{bmatrix} -A^{(1)} \\ A^{(2)} \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

yields the same discriminant function as the formulation

$$\min_{w} ||A_2 w - B|| = \min_{w^{(1)}} ||A_2 w^{(1)} - b_1|| + \min_{w^{(2)}} ||A_2 w^{(2)} - b_2||$$

where

$$A_2 = \begin{bmatrix} A^{(1)} \\ A^{(2)} \end{bmatrix} \qquad w = \begin{bmatrix} w^{(1)} & w^{(2)} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} b_1 & b_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}0\\\vdots\\0\end{matrix}}\right\} N_1 \\ \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} N_2 \end{matrix}$$

$A^{(1)}$, $A^{(2)}$, N_1, N_2 are as defined in Ch. I. For the first formulation the solution is

$$\hat{w} = A_1^{\dagger} b$$

For the second formulation we get:

$$\hat{w}^{(1)} = A_2^{\dagger} b_1, \quad \hat{w}^{(2)} = A_2^{\dagger} b_2$$

the corresponding weight vector for the resulting hyperplane is

$$\hat{w}^{(1)} - \hat{w}^{(2)} = A_2^{\dagger} (b_1 - b_2) = A_2^{\dagger} \begin{bmatrix} -1 \\ \vdots \\ -1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{matrix} \} N_1 \\ \\ \} N_2 \end{matrix}$$

but this is a solution to the problem

$$A_2 w = \begin{bmatrix} -1 \\ \vdots \\ -1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{matrix} \} N_1 \\ \\ \} N_2 \end{matrix}$$

which is the same as:

$$A_1 w = b = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

hence the two formulations yield the same discriminant function.

Let us look at the formulation $||A_1 w - b||^2$. Repeating a pattern that is misclassified corresponds by theorem 3 to changing b_i to $b_i - \rho_K (a_i^* \hat{w}_K - b_i)$. Here we have $b_i = 1$, $0 < \rho_K < 1$, $a_i^* \hat{w}_K < 0$ if the pattern is misclassified. Hence $-\rho_K (a_K^* \hat{w}_K - b_i) > 0$. This corresponds in the matrix B to increasing the cost of misclassification.

The procedure had been applied to Sebestyen and Edie's data [54] consisting of 168 two-dimensional vectors representing six classes,

Fig. 3. All the patterns that were in error after the MSE solution were repeated. The corresponding new solution vector is found and the resulting misclassified patterns were repeated. This was carried on several times for different numbers of features. The results of these iterations are shown in the following pages for 2, 3, and 5 unaugmented features. The results are tabulated in Table 3 and compare favorably with the MSE solution. Table 2 gives the results of the MSE solution up to 33 features. It is noted that we get fewer errors using 5 features than the MSE solution using 33 features. We also note that the number of misclassifications are almost the same for 2, 3, and 5 features. In the subsequent parts of this dissertation calculations will be carried out only on 2 and 3 unaugmented features.

Table 2. MSE Results [68]

	Order of the Polynomial									
	1		2		3		4		5	
Pat-	2 Features		5 Features		9 Features		17 Features		33 Features	
tern Class	No.* Misc.	%** Misc.	No. Misc.	% Misc.	No. Misc.	% Misc.	No. Misc.	% Misc.	No. Misc.	% Misc.
1	6	100	6	100	4	66.67	2	33.33	1	16.67
2	0	6	0	0	4	10.25	3	7.80	3	7.80
3	0	0	2	12.50	1	6.25	1	6.25	1	6.25
4	0	0	3	8.10	1	2.70	3	8.10	2	5.40
5	18	78.25	3	13.05	3	13.05	1	4.35	2	8.70
6	33	70.25	6	12.80	3	6.38	3	6.38	3	6.38
Total	57	33.95	20	11.90	16	9.52	13	7.74	12	7.14
	%Rec. = 66.05		%Rec. = 88.10		% Rec. = 90.48		% Rec. = 92.26		%Rec. = 92.86	

*No. Misc. = Number of Misclassification.
** % Misc. = Percent of Misclassification.
% Rec. = Percent of Correct Classification.

Note that for the polynomial of order 1 the features are x_1 and x_2, for the polynomial of order 2 the features are x_1, x_2, x_1^2, x_1x_2, x_2^2; that is, the features are x_1, x_2 plus those features generated from $(x_1 + x_2)^2$. For the third polynomial the features are x_1, x_2, x_1^2, $x_1\ x_2$, x_2^2 plus those features generated from $(x_1 + x_2)^3$ ...

Table 3. Comparison of the Number of Misclassifications Resulting From Repeating the Samples That Are in Error With the MSE Results.

	No. of Misclassifications	% of Misclassifications
MSE Solution 2 features	57	33.95
Iterative Weighting 2 features	16	9.5
MSE Solution 3 features	40	23.8
Iterative Weighting 3 features	15	9
MSE Solution 5 features	20	11.9
Iterative Weighting 5 features	10	5.95

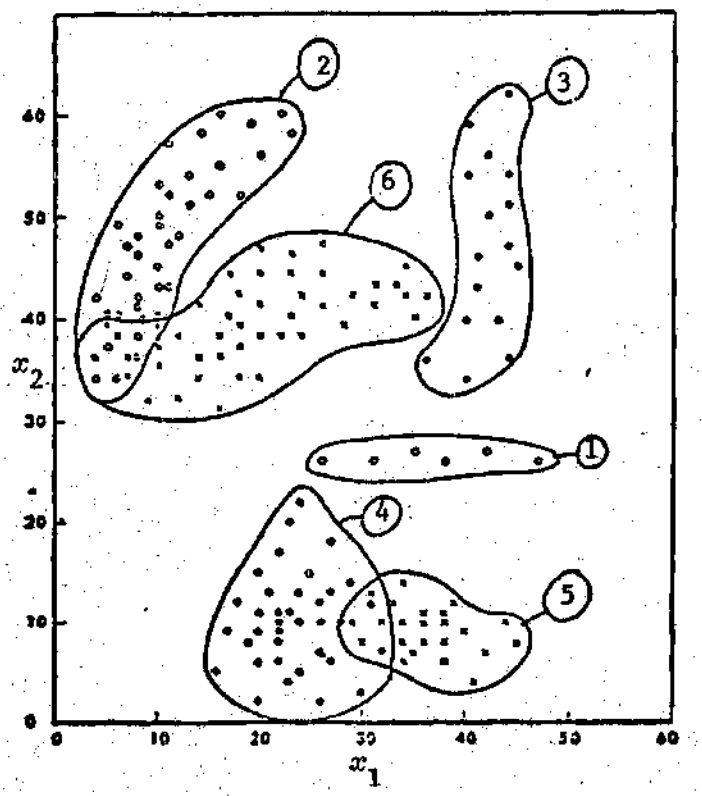


Figure 3. 168 Two-Dimensional Vectors

Repeating all the patterns that are in error

2 features (x_1, x_2)

Class	MSE No. Misc.	1st Iteration	2nd Iteration	3rd Iteration	4th Iteration
1	6	6	6	6	6
2	0	39	0	18	0
3	0	2	0	1	1
4	0	10	0	3	0
5	18	0	4	2	2
6	33	0	38	3	20
	57	57	48	33	29

Class	5th Iteration	6th Iteration	7th Iteration	8th Iteration	9th Iteration	10th Iteration
1	4	4	3	2	2	2
2	22	0	14	0	7	3
3	5	2	2	2	1	1
4	3	2	2	2	2	2
5	1	2	2	2	2	2
6	0	25	1	13	4	6
	35	35	24	21	18	16

Repeating all the Patterns that are in error

3 features (x_1, x_2, x_1x_2)

Class	MSE No. Misc.	1st Iteration	2nd Iteration	3rd Iteration	4th Iteration
1	6	6	6	6	5
2	0	15	4	4	4
3	0	2	0	1	1
4	2	2	2	1	2
5	2	2	2	3	2
6	30	0	8	4	5
	40	27	22	19	19

Class	5th Iteration	6th Iteration	7th Iteration	8th Iteration	9th Iteration	10th Iteration
1	3	2	2	2	2	2
2	4	5	4	5	4	4
3	1	1	1	1	2	1
4	2	2	2	2	2	2
5	2	2	2	2	2	2
6	6	6	6	6	6	4
	18	18	17	18	18	15

Repeating all the patterns that are in error

5 features $(x_1, x_2, x_1^2, x_1x_2, x_2^2)$

Class	MSE No. Misc.	1st Iteration	2nd Iteration	3rd Iteration	4th Iteration
1	6	6	4	2	2
2	0	4	4	4	4
3	2	1	1	2	1
4	3	3	2	2	2
5	3	3	3	3	2
6	6	3	3	4	5
	20	17	17	17	17

Class	5th Iteration	6th Iteration	7th Iteration	8th Iteration	9th Iteration	10th Iteration
1	3	2	2	2	2	1
2	2	4	3	4	2	4
3	1	2	2	2	1	2
4	2	2	2	2	2	2
5	2	2	2	2	2	2
6	5	5	5	5	5	3
	15	17	16	17	14	14

Class	11th Iteration	12th Iteration	13th Iteration	14th Iteration	15th Iteration	16th Iteration
1	1	2	1	1	2	1
2	3	3	3	3	4	2
3	2	2	2	1	1	1
4	3	2	2	2	2	2
5	2	2	2	2	2	2
6	3	4	4	4	4	4
	14	15	14	13	15	12

Class	17th Iteration	18th Iteration	19th Iteration	20th Iteration	21st Iteration	22nd Iteration
1	0	0	2	1	0	1
2	4	2	5	1	2	3
3	1	1	1	1	1	0
4	2	2	2	2	2	2
5	2	2	2	2	2	2
6	4	4	4	6	4	4
	13	11	16	13	11	12

Class	23rd Iteration	24th Iteration	25th Iteration
1	0	0	0
2	2	4	2
3	1	0	0
4	2	2	2
5	2	2	2
6	4	4	4
	11	12	10

CHAPTER III

A NEW ALGORITHM FOR PATTERN CLASSIFICATION

In this chapter the algorithm resulting from the method of repeating misclassified samples is presented. The algorithm is versatile in the sense that it can take any of three different forms due to the theorems presented in Chapter II. One form might be more attractive computationally for a particular problem than another. The algorithm is compared with the relaxation and the Ho-Kashyap algorithms. The algorithm yields a better solution than the MSE solution in the non-separable case. A convergence proof to a separating solution in the linearly separable case is presented in Theorem 4.

The Algorithm

The algorithm takes the form of checking the samples for correct classifications and repeating the misclassified samples. The construction for the two-class problem is the same as in Lemma 1. For single pattern adaptation the algorithm takes the form:

$$w_{K+1} = \begin{cases} w_K - \rho_K \, (A_K^* A_K)^\dagger \, a_i \, (a_i^* w_K - b_i), & \text{if } a_i^* w_K \leq 0 \\ w_K \quad \text{if } a_i^* w_K > 0 \end{cases}$$

where $$\rho_K = \frac{1}{1 + a_i^* \, (A_K^* A_K)^\dagger \, a_i}$$

The algorithm can take the following three forms due to the three theorems presented in Chapter II. The K-subscript is updated only if w_K is changed. The algorithm starts with the MSE solution of the original system where

$$A = \begin{bmatrix} \\ a_1^* \\ \vdots \\ a_i^* \\ \vdots \\ a_m^* \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} .$$

Form I

In this form A_K is changing while b_i is kept constant and will be taken as 1. The algorithm takes the form,

$$w_{K+1} = \begin{cases} w_K - \rho_K \, (A_K^* A_K)^\dagger \, a_i \, (a_i^* w_K - 1), & \text{if } a_i^* w_K \leq 0 \\ w_K & \text{if } a_i^* w_K > 0 \end{cases}$$

where ρ_K is defined as before, and $w_o = A_o^\dagger b$ where $A_o = A$.

Form II

This form results from Theorem 2. The algorithm takes the form,

$$w_{K+1} = \begin{cases} A_{K+1}^\dagger \, b_{K+1} & \text{if } a_i^* w_K \leq 0 \\ w_K & \text{if } a_i^* w_K > 0 \end{cases}$$

where

$$A_K = \begin{bmatrix} \\ a_{K1}^* \\ \vdots \\ a_{Ki}^* \\ \vdots \\ a_{Km}^* \end{bmatrix} , \qquad b_K = \begin{bmatrix} b_{K1} \\ \vdots \\ b_{Ki} \\ \vdots \\ b_{Km} \end{bmatrix} .$$

$$A_{K+1} = \begin{bmatrix} a_{K1}^* \\ \vdots \\ \sqrt{2}\, a_{Ki}^* \\ \vdots \\ a_{Km}^* \end{bmatrix}, \qquad b_{K+1} = \begin{bmatrix} b_{K1} \\ \vdots \\ \sqrt{2}\, b_{Ki} \\ \vdots \\ b_{Km} \end{bmatrix}$$

and $$A_o = A = \begin{bmatrix} a_1^* \\ \vdots \\ a_i^* \\ \vdots \\ a_m^* \end{bmatrix}, \qquad b_o = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Form III

This form results from Theorem 3. The matrix is kept unchanged while the vector b is updated. The algorithm takes the form,

$$w_{K+1} = \begin{cases} w_K + \rho_K (A^*A)^\dagger a_i (b_{Ki} - a_i^* w_K) = A^\dagger b_{K+1}, & \text{if } a_i^* w_K \leq 0 \\ w_K, & \text{if } a_i^* w_K > 0 \end{cases}$$

or,

$$w_{K+1} = \begin{cases} (A^*A)^\dagger A^* \begin{bmatrix} b_{K1} \\ \vdots \\ b_{Ki} + \rho_K (b_{Ki} - a_i^* w_K) \\ \vdots \\ b_{Km} \end{bmatrix}, & \text{if } a_i^* w_K \leq 0 \\ w_K, & \text{if } a_i^* w_K > 0 \end{cases}$$

where

$$b_{K+1} = \begin{bmatrix} b_{K1} \\ \vdots \\ b_{Ki} + \rho_K (b_{Ki} - a_i^* w_K) \\ \vdots \\ b_{Km} \end{bmatrix}$$

$$b_K = \begin{bmatrix} b_{K1} \\ \vdots \\ b_{Ki} \\ \vdots \\ b_{Km} \end{bmatrix} \qquad \text{and} \qquad b_o = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}$$

and where

$$\rho_K = \frac{1}{1 + a_i^* (A^* A)^\dagger a_i}$$

This form might be more attractive computationally since only the generalized inverse $(A^* A)^\dagger$ is required.

Comparison With Other Algorithms

a) The relaxation algorithm

The single pattern adaptation of the relaxation algorithm is,

$$w_{K+1} = \begin{cases} w_K + \rho \dfrac{b - a_i^* w_K}{||a_i||^2} a_i , & \text{if } a_i^* w_K \leq b \\ w_K , & \text{if } a_i^* w_K > b \end{cases}$$

where $0 < \rho < 2, b > 0$

After a correction we have

$$(a_i^* w_{K+1} - b) = (1 - \rho)(a_i^* w_K - b)$$

If $\rho < 1$, we have under-relaxation. If $\rho > 1$ we have over-relaxation. Comparing this with the results of Theorem 1, we get after correction for the new algorithm

$$(a_i^* w_{K+1} - b) = (1 - a_i^* h_{K+1})(a_i^* w_K - b)$$

where

$$a_i^* h_{K+1} = \frac{a_i^* (A_K^* A_K)^\dagger a_i}{1 + a_i^* (A_K^* A_K)^\dagger a_i} < 1$$

which makes the new algorithm analogous to under-relaxation.

b) The Ho-Kashyap Algorithm

The perceptron and relaxation procedures find separating vectors if the samples are linearly separable, but do not converge on nonseparable problems. The Ho-Kashyap procedures, being a MSE procedure, yields a weight vector whether the patterns are linearly separable or not. If the patterns are linearly separable, the Ho-Kashyap algorithm yields a separating vector in a finite number of steps. If the patterns are not linearly separable, it provides us with evidence of nonseparability. However, there is no bound on the number of steps needed to disclose nonseparability. The method for finding the weight vector involves the minimization of J(w,b) with respect to w and b.

$$J(w, b) = \frac{1}{2}||Aw - b||^2 , \quad b > 0 .$$

A gradient method for minimizing J(w, b) can be developed by changing w and b alternately in the direction of the negative gradients with respect to w and b respectively where

$$\frac{\partial J}{\partial w} = A^T (Aw - b)$$

$$\frac{\partial J}{\partial b} = (b - Aw) .$$

Since w is not constrained, $\frac{\partial J}{\partial w} = 0$ implies $w = A^\dagger b$.

Next b is changed in the direction of the negative of the gradient with

respect to b under the constraint $b > 0$. This means that only those components of b such that changing them in the direction of the negative of the gradient with respect to b will become more positive will be changed. The initial value of b is taken to be

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} .$$

Thus the Ho-Kashyap algorithm takes the form,

$$w_K = A^{\dagger} b_K , \qquad b_o^T = [1, 1, \ldots, 1]$$

$$w_{K+1} = \begin{cases} w_K + \rho (A^*A)^{\dagger} a_i (a_i^* w_K - b_{Ki}) = A^{\dagger} b_{K+1} ; & \text{if } a_i^* w_K > b_{Ki} \\ w_K \qquad \text{if } a_i^* w_K \leq b_{Ki} \end{cases}$$

where

$$b_K = \begin{bmatrix} b_{K1} \\ \vdots \\ b_{Ki} \\ \vdots \\ b_{Km} \end{bmatrix} \qquad \text{and} \qquad b_{K+1} = \begin{bmatrix} b_{K1} \\ \vdots \\ b_{Ki} + \rho (a_i^* w_K - b_{Ki}) \\ \vdots \\ b_{Km} \end{bmatrix}$$

Hence, the components of b corresponding to the samples that are correctly classified are changed. The procedure converges to a separating solution in the linearly separable case if $0 < \rho \leq 1$. Ho and Kashyap observe that convergence can be improved by varying ρ at every stage, but they offer no suggestion on how to make these variations. The procedure presented in this dissertation bears a great deal of similarities to the Ho-Kashyap algorithm. The procedure that we

presented starts with the MSE solution and then repeats those samples that are in error. This was shown by theorem 3 to be equivalent to changing b_{Ki} in the direction of the gradient with respect to b. This means changing the components of b that are in error such that the corresponding b_{Ki} is increased.

Both the relaxation and the Ho-Kashyap algorithms yields a separating solution in the linearly separable case. So does the new algorithm. The Ho-Kashyap algorithm gives an indication of nonseparability at any stage when $(Aw_K - b_K) < 0$, [25], but so does the new algorithm since it is also a MSE procedure. In the nonseparable case, the Ho-Kashyap algorithm has the advantage over the relaxation algorithm of reverting to the original MSE solution [28]. The greatest advantage of the new algorithm, beside its being statistically more appealing than the Ho-Kashyap algorithm, is that we can do better than the original MSE solution.

Convergence Proof

Theorem 4. Let S_1 and S_2 be two classes of linearly separable prototypes. The new algorithm converges to a solution that separates S_1 and S_2 in a finite number of steps.

Proof. In the proof we will keep A constant and change b (see Theorem 3). The algorithm takes the form

$$w_{K+1} = w_K + \rho_K (A^*A)^\dagger a_i (b_{Ki} - a_i^* w_K) \; ; \quad a_i^* w_K \leq 0$$

multiplying the above equation by A yields

$$Aw_{K+1} = Aw_K + \rho_K A(A^*A)^\dagger a_i (b_{Ki} - a_i^* w_K) .$$

If the samples are linearly separable then there exists a solution vector $\hat{w}$ such that $A\hat{w} = \hat{b} > 0$. Choose $\alpha > 0$ such that $\alpha\, a_i^* \hat{w} > b_{Ki}$ for all K and i. Clearly $\alpha\hat{w}$ is also a solution since $\alpha A\hat{w} > 0$. Substracting $\alpha A\hat{w}$ from both sides of the previous equation yields,

$$Aw_{K+1} - \alpha A\hat{w} = Aw_K - \alpha A\hat{w} + \rho_K\, A(A^*A)^\dagger a_i\, (b_{Ki} - a_i^* w_K) .$$

Hence,

$$||Aw_{K+1} - \alpha A\hat{w}||^2 = ||Aw_K - \alpha A\hat{w}||^2$$

$$+ 2\rho_K\, a_i^* (A^*A)^\dagger A^*(Aw_K - \alpha A\hat{w})\, (b_{Ki} - a_i^* w_K)$$

$$+ ||\rho_K\, A(A^*A)^\dagger a_i\, (b_{Ki} - a_i^* w_K)||^2 \quad .$$

The second and third terms on the right side of the above equation simplify considerably. The second term simplifies by noting that $(A^*A)^\dagger A^* = A^\dagger$ and that $A^*A\, w_K = A^*\, b_K$. Thus the second term becomes

$$2\rho_K\, (b_{Ki} - a_i^* w_K)\, a_i^*\, (A^*A)^\dagger A^*(A\, w_K - \alpha A\hat{w})$$

$$= 2\rho_K\, (b_{Ki} - a_i^* w_K)\, a_i^*\, (A^*A)^\dagger A^*(b_K - \alpha\hat{b})$$

$$= 2\rho_K\, (b_{Ki} - a_i^* w_K)\, a_i^*\, A^\dagger\, (b_K - \alpha\hat{b})$$

$$= 2\rho_K\, (b_{Ki} - a_i^* w_K)\, a_i^*\, (w_K - \alpha\hat{w})$$

$$\leq - 2\rho_K\, (b_{Ki} - a_i^* w_K)^2$$

since $\alpha\, a_i^* \hat{w} > b_{Ki}$.

The third term simplifies by noting that

$$|| \rho_K A(A^*A)^\dagger a_i (b_{Ki} - a_i^* w_K)||^2 = ||A(w_{K+1} - w_K)||^2$$

$$= ||AA^\dagger b_{K+1} - AA^\dagger b_K||^2 = ||AA^\dagger (b_{K+1} - b_K)||^2$$

$$= ||P_{R(A)} (b_{K+1} - b_K)||^2$$

$$\leq ||b_{K+1} - b_K||^2 = \rho_K^2 (b_{Ki} - a_i^* w_K)^2$$

since $AA^\dagger$ is an orthogonal projector on the range of A. Thus we have,

$$||Aw_{K+1} - \alpha A\hat{w}||^2 \leq ||Aw_K - \alpha A\hat{w}||^2 - \rho_K (2 - \rho_K)(b_{Ki} - a_i^* w_K)^2 \quad .$$

Since $\rho_K < 1$, we have

$$\rho_K (2 - \rho_K)(b_{Ki} - a_i^* w_K)^2 > \rho_K (b_{Ki} - a_i^* w_K)^2 \quad .$$

Let $\delta^2 = \min_{K,i} \rho_K (b_{Ki} - a_i^* w_K)$

then

$$||Aw_{K+1} - \alpha A\hat{w}||^2 < ||Aw_K - \alpha A\hat{w}||^2 - \delta^2 \quad .$$

Taking $W_o = A^\dagger b_o$, where $b_o = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

we get after K corrections,

$$||Aw_K - \alpha A\hat{w}|| < ||Aw_o - \alpha A\hat{w}||^2 - K\delta^2 \quad .$$

Since the squared distance cannot become negative, it follows that the sequence of corrections must terminate after no more than K_o correction,

where

$$K_o = \frac{||Aw_o - \alpha A\hat{w}||^2}{\delta^2} = \frac{||b_o - \alpha\hat{b}||^2}{\delta^2}$$

Thus when correction ceases the resulting weight vector must classify all the samples correctly.

Examples

Example 1. The algorithm was carried out on a 4-variable separable problem for which the solution is obtained on the first iteration. In a switching problem with n-variables [24], [12], one is concerned with the vertices of an n-cube, each one of which may be assigned to only one of two classes, A and B. It is required to find a separating hyperplane if one exists. An (n + 1) vector is associated with every vertex where the additional components is always $\pm$ 1, the augmented component. The n components are the coordinates of the vertex in n-dimensional space. The values of the components take on the values $\pm$ 1 only. An ordering of the 2^n vertices is formed by n-variables in natural binary code [12]. Thus, any vertex can be identified by the decimal number corresponding to the input binary work. Thus the decimal number 6 corresponds to the vertex: -1 1 1 -1 and the corresponding augmented vector (-1, 1, 1, -1, 1). The example is example 2 in Ho and Kashyap's paper [24]. It consists of 16 separable vertices, with eight vertices in each class.

Class A = {7, 9 to 15}
Class B = {0 to 6, 8 }

Only one iteration of the two samples 7 and 8 was needed to get a separable solution, Ho and Kashyap obtained a solution on the first iteration.

Example 2. For the separable subset of Sebestyen and Edie's data consisting of classes 2, 3, and 4, a separating solution was obtained on the zeroth iteration.

Example 3. For the separable subset of Sebestyen and Edie's data consisting of classes 1, 5, and 6, a separating solution was obtained on the 22nd iteration.

CHAPTER IV

APPLICATION OF CONSTRAINED GENERALIZED INVERSE TO PATTERN CLASSIFICATION

In this chapter an adaptive constraint procedure is presented and applied to Sebestyen and Edie's data with very favorable results. The procedure utilizes the means of the different classes. The MSE solution results in weight vectors proportional to the means of the classes and so does a broad class of criterion functions.

The proposed method to combat MSE deficiencies and achieve smaller percentage of misclassification is to introduce linear constraints. The problem takes the form:

$$\min_{W} \; ||AW - B||^2 \quad \text{subject to } M^T W = F$$

where M^T and F are given matrices of appropriate dimensions. The generalized inverse gives the minimum norm solution.

The generalized inverse solution is [36]

$$\hat{W}^* = A^\dagger [I - (M^T A^\dagger)^\dagger \; M^T A^\dagger] \; B + A^\dagger \; (M^T A^\dagger)^\dagger \; F \quad .$$

$$= A^\dagger \{B - (M^T A^\dagger)^\dagger \; [M^T A^\dagger B - F]\} \quad .$$

If A is of full-column rank and M^T has a full row rank, then all the inverses exist and we can get the following solution by the Lagrange multiplier method.

$$\hat{W}^* = (A^TA)^{-1} \{A^TB - M [M^T(A^TA)^{-1}M]^{-1} [M^T(A^TA)^{-1} A^TB - F]\}$$

The solution without constraint is:

$$\hat{W} = A^{\dagger}B$$

which becomes when A is of full column rank

$$\hat{W} = (A^TA)^{-1} A^TB \quad .$$

Expressing $\hat{W}^*$ in terms of $\hat{W}$ we get,

$$\hat{W}^* = \hat{W} - A^{\dagger}(M^TA^{\dagger})^{\dagger} [M^T \hat{W} - F]$$

which for the case when A is of full column rank and M^T of full row rank takes the form

$$\hat{W}^* = \hat{W} - (A^TA)^{-1} M[M^T(A^TA)^{-1} M]^{-1} [M^T\hat{W} - F] \quad .$$

Utilizing the information that we get from the MSE solution, we can impose appropriate linear constraints. Note that we already have $\hat{W}$ and $(A^TA)^{-1}$ from the unconstrained solution, so computationally the above formulation is attractive. Hence, we suggest an adaptive constraint setting motivated by Fisher linear discriminant. Koford and Groner [30] showed the the MSE solution is equal to Fisher's for appropriate choice of B. Peterson and Mattson [49] and Smith [58] suggested a parametric search for the matrix B. Smith [59] raised the question of whether the different adaptive algorithms might be improved by suitable constraint, but observed that the necessary method of constraint is not obvious.

Fisher linear discriminant (for 2 classes) projects the data onto a line and chooses the line that makes the difference between the projected means of each class as large as possible (Fig. 4). For K

classes, the projection is taken onto a K-1 dimensional space.

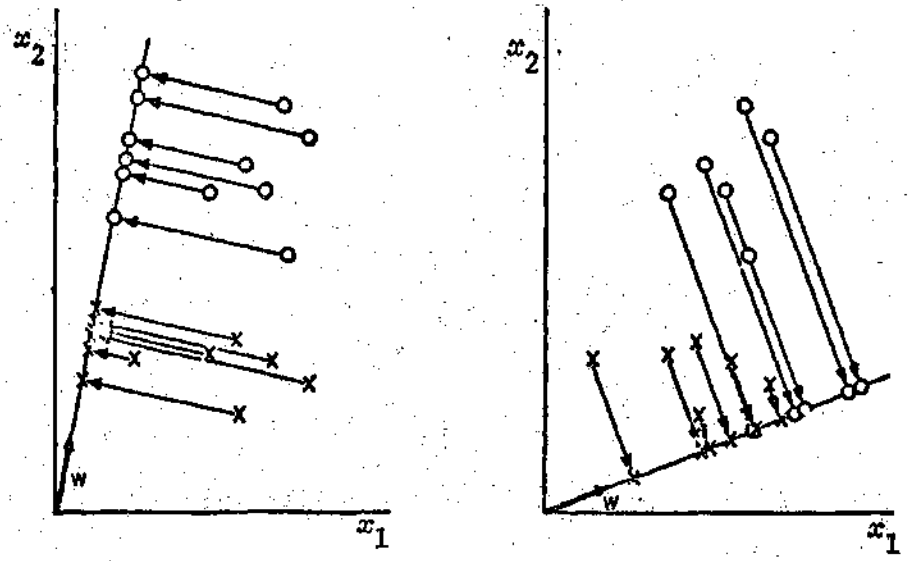


Figure 4. Projection of Samples Onto a Line.

Maximizing the projected-means difference corresponds to maximizing $|w^T(m_1 - m_2)|$ where m_1 is the mean of the first class and m_2 is the mean of the second class. In our terminology $w = w^{(1)} - w^{(2)}$ and we are dealing with the quantities $w^{(1)T}m_1$, $w^{(1)T}m_2$, $w^{(2)T}m_1$, $w^{(2)T}m_2$.

When $x^{(i)}$ is restricted to multivariate Gaussian with mean m_i and covariance $\Phi^{(i)}$ and if $\Phi^{(1)} = \Phi^{(2)}$, then w is proportional to $[\Phi^{(1)} + \Phi^{(2)}]^{-1}(m_1 - m_2)$ and the threshold weight $\omega_p = \frac{1}{2}(m_1 + m_2)^T w$, together yielding the classifier obtained in [30]. This is minimum error optimal for Gaussian equal covariance distributions with equal costs. Note that $\omega_p = -\frac{1}{2}(m_1 + m_2)^T w$ is proportional to $(m_1 + m_2)^T [\Phi^{(1)} + \Phi^{(2)}]^{-1}(m_1 - m_2)$. Thus the threshold weight vector, ω_p, is also proportional to the difference between the means.

Peterson and Mattson [49] showed that the weight vector w is

proportional to the difference of the means for a broader class of criterion functions, any differentiable function of the means and the variances of the classes, and showed that w satisfies an equation of the form:

$$[\alpha_1 \Phi^{(1)} + \alpha_2 \Phi^{(2)}] w = (m_1 - m_2) \alpha_3$$

where α_1, α_2 and α_3 are scalars. They suggested a parametric search for α_1, α_2 and α_3.

Smith [58] observed that since the misclassification usually occurs at the tails of distributions, any distribution with Gaussian tails could be assumed to be Gaussian for correction purposes and proposed a parametric search procedure for the ratio of costs vectors.

The MSE approach yields veight vectors proportional to the means of the classes. More specifically:

$$\hat{W} = A^{\dagger}B = (A^TA)^{-1} A^T B \ , \quad \gamma(j/i) = \begin{cases} 0 & \text{if } i = j \\ \gamma > 0 & \text{otherwise} \end{cases}$$

$$\hat{W} = (A^TA)^{-1} [\gamma_{i\neq 1} \Sigma N_i m_i \quad \gamma_{i\neq 2} \Sigma N_i m_i \ \ldots\ \gamma_{i\neq k} \Sigma N_i m_i]$$

$$\hat{W} = \gamma N (A^TA)^{-1} \left[-\frac{N_1}{N} m_1 + m \quad -\frac{N_2}{N} m_2 + m \ \ldots\ -\frac{N_K}{N} m_K + m\right]$$

where m is the mean of all the samples. Since our decision rule is $w^{(i)^T} x < w^{(j)^T} x$, $x \in \omega_i$; γ, N and m, being common to all terms, can be dropped, and we get

$$\hat{W}^{(i)} = (A^TA)^{-1} \left(-\frac{N_i}{N} m_i\right)$$

A constraint usually utilizes some a priori information. Clearly after

we get the MSE solution and test it on the patterns we have more information about its goodness. By utilizing the means of the classes we could change the position of the resulting hyperplanes so that they are closer or further away from the mean of a particular class.

Whether the constrained MSE is superior or inferior depends on the criterion chosen and on the difference between the correct constraint and the imposed constraint. In this dissertation the criterion is minimizing the error on the design set.

The Proposed Constraint

$$M^T \hat{W} = F$$

$$M^T = \begin{bmatrix} m_1^T \\ \vdots \\ m_6^T \end{bmatrix} \quad ; \quad \hat{W} = [\hat{w}^{(1)} \ \ldots \ \hat{w}^{(6)}]$$

$$M^T \hat{W} = \begin{bmatrix} m_1^T \hat{w}^{(1)} & m_1^T \hat{w}^{(2)} & m_1^T \hat{w}^{(3)} & m_1^T \hat{w}^{(4)} & m_1^T \hat{w}^{(5)} & m_1^T \hat{w}^{(6)} \\ m_2^T \hat{w}^{(1)} & \cdots & \cdots & \cdots & \cdots & m_2^T \hat{w}^{(6)} \\ \vdots & & & & & \\ m_6^T \hat{w}^{(1)} & \cdots & \cdots & \cdots & \cdots & m_6^T \hat{w}^{(6)} \end{bmatrix}$$

Procedure

1) Obtain $m^T \hat{W}$ and the number of errors, and with which class

the errors are committed.

2) Start with the class that is most misclassified and look at $m_i^T \hat{w}^{(i)}$ corresponding to that class. If the mean of the class is getting misclassified or getting classified correctly but with small margin, put the constraint on $m_i^T \hat{w}^{(i)}$ such that its $m_i^T \hat{w}^{(i)*}$ is less than $m_i^T \hat{w}^{(i)}$. (Remember our decision rule is if $x^T \hat{w}^{(i)} < x^T \hat{w}^{(j)}$ then x is classified in class i.) If on the other hand $m_i^T \hat{w}^{(i)}$ is very small, then decreasing it will not improve the situation, but look at the class with which the error is committed, say ω_j, and decrease its $m_j^T \hat{w}^{(j)}$. Note that by working with the diagonal elements, we are only affecting that particular weight vector while the others are not changing.

3) Adjust the diagonal elements of the class that has the greatest number of misclassified samples and the class where these samples are being classified until you get the best possible result without affecting adversely the other classes.

4) Carry on procedure 3) between two classes at a time, moving to the class that has the next largest number of misclassifications.

5) After getting the best possible adjustments with the diagonal element, go the off-diagonal elements.

6) A finer refinement is possible by decreasing the change in adjustment.

7) Take the change in adjustment appropriately in our case we used a change of 10 when 1000 was used for B. We could get more quantitative and choose:

$$\frac{\min_i \left| m_i^T \hat{w}^{(i)} \right|}{\max_{i,j} \left| m_{ij} \right|} \Delta x$$

where m_{ij} is the j^{th} coordinate value of the i^{th} class mean and Δx is the smallest distance between samples or the average distance between samples. The step size of 10 corresponds to Δx being the smallest distance between samples in our example.

The above procedure was carried on Sebestyen and Edie's data [54], Fig. 3. The results of the MSE solution are shown below for the unconstrained and the constrained case with the corresponding $M^T\hat{W} - F$ matrix. Table 4 compares the results of the MSE solution with and without constraint.

In the biased means we shifted the means a bit to the tails of the the distributions by dividing by $N_i - 1$ instead of N_i, where N_i is the number of samples in class i. This was done for all classes except class 1 which has the smallest number of samples and all of its members were misclassified.

Computer Results

2 features (x_1, x_2)

a) No Constraints

NO.OF MISCLASSIFICATIONS IN CLASS 1= 6

NO.OF MISCLASSIFICATIONS IN CLASS 2= 0

NO.OF MISCLASSIFICATIONS IN CLASS 3= 0

NO.OF MISCLASSIFICATIONS IN CLASS 4= 0

NO.OF MISCLASSIFICATIONS IN CLASS 5= 18

NO.OF MISCLASSIFICATIONS IN CLASS 6= 33

$$M^T\hat{W} - F = \begin{bmatrix} .0000 & .0000 & .0000 & .0000 & .0000 & .0000 \\ .0000 & .0000 & .0000 & .0000 & .0000 & .0000 \\ .0000 & .0000 & .0000 & .0000 & .0000 & .0000 \\ .0000 & .0000 & .0000 & .0000 & .0000 & .0000 \\ .0000 & .0000 & .0000 & .0000 & .0000 & .0000 \\ .0000 & .0000 & .0000 & .0000 & .0000 & .0000 \end{bmatrix}$$

Table 4. Constrainted MSE Results.

	No. of Misclassifications	% of Misclassification
MSE Solution 2 features	57	33.95
MSE With Constrained Biased Means, 2 features	16	9.5
MSE With Constrained Means, 2 features	24	14.3
MSE Solution 3 features	40	23.8
MSE With Constrained Biased Means, 3 features	22	13.1
MSE With Constrained Means, 3 features	20	11.9

b) With Constraints (biased means)

```
NO.OF MISCLASSIFICATIONS IN CLASS  1=          0
NO.OF MISCLASSIFICATIONS IN CLASS  2=          6
NO.OF MISCLASSIFICATIONS IN CLASS  3=          0
NO.OF MISCLASSIFICATIONS IN CLASS  4=          3
NO.OF MISCLASSIFICATIONS IN CLASS  5=          0
NO.OF MISCLASSIFICATIONS IN CLASS  6=          7
```

$$
\begin{bmatrix}
 .1200+03 & .0000 & .0000 & .0000 & .0000 & .0000 \\
 .0000 & -.3000+02 & .0000 & .0000 & .0000 & .0000 \\
 .0000 & .0000 & -.3000+02 & .0000 & .0000 & .0000 \\
 .0000 & .0000 & .0000 & -.5000+02 & .0000 & .0000 \\
 -.2600+03 & .0000 & .0000 & .0000 & -.3700+02 & .0000 \\
 .0000 & .0000 & .0000 & .0000 & .0000 & .3000+02
\end{bmatrix} M^T W - F
$$

c) With Constraints (means)

```
NO.OF MISCLASSIFICATIONS IN CLASS  1=          3
NO.OF MISCLASSIFICATIONS IN CLASS  2=          9
NO.OF MISCLASSIFICATIONS IN CLASS  3=          0
NO.OF MISCLASSIFICATIONS IN CLASS  4=          2
NO.OF MISCLASSIFICATIONS IN CLASS  5=          3
NO.OF MISCLASSIFICATIONS IN CLASS  6=          7
```

$$
\begin{bmatrix}
 .4000+02 & .0000 & .0000 & .0000 & .0000 & .0000 \\
 .0000 & .4000+02 & .0000 & .0000 & .0000 & .0000 \\
 .0000 & .0000 & .2000+02 & .0000 & .0000 & .0000 \\
 .0000 & .0000 & .0000 & .3500+02 & .0000 & .0000 \\
 -.4000+02 & .0000 & .0000 & .0000 & -.2000+02 & .0000 \\
 .0000 & .0000 & .0000 & .0000 & .0000 & .4000+02
\end{bmatrix} M^T Q - F
$$

3 features

a) No Constraints

NO.OF MISCLASSIFICATIONS IN CLASS 1= 6

NO.OF MISCLASSIFICATIONS IN CLASS 2= 0

NO.OF MISCLASSIFICATIONS IN CLASS 3= 0

NO.OF MISCLASSIFICATIONS IN CLASS 4= 2

NO.OF MISCLASSIFICATIONS IN CLASS 5= 2

NO.OF MISCLASSIFICATIONS IN CLASS 6= 30

.0000	.0000	.0000	.0000	.0000	.0000
.0000	.0000	.0000	.0000	.0000	.0000
.0000	.0000	.0000	.0000	.0000	.0000
.0000	.0000	.0000	.0000	.0000	.0000
.0000	.0000	.0000	.0000	.0000	.0000
.0000	.0000	.0000	.0000	.0000	.0000
.9125+03	.9824+03	.7578+03	.8755+03	.6367+03	.8352+03
.9936+03	.2589+03	.1055+04	.1141+04	.9644+03	.5873+03
.8939+03	.1087+04	.3089+03	.1098+04	.9744+03	.5375+03
.9754+03	.1074+04	.9892+03	.3187+03	.8140+03	.8284+03
.8946+03	.9591+03	.1040+04	.8376+03	.2151+03	.1054+04
.9755+03	.6330+03	.8732+03	.8976+03	.9898+03	.6310+03

$M^T\hat{W} = F$

$M^T\hat{W}$

b) With Constraints (biased means)

NO.OF MISCLASSIFICATIONS IN CLASS 1= 6

NO.OF MISCLASSIFICATIONS IN CLASS 2= 4

NO.OF MISCLASSIFICATIONS IN CLASS 3= 0

NO.OF MISCLASSIFICATIONS IN CLASS 4= 2

NO.OF MISCLASSIFICATIONS IN CLASS 5= 2

NO.OF MISCLASSIFICATIONS IN CLASS 6= 8

.0000	.0000	.0000	.0000	.0000	.0000
.0000	-.5000+02	.0000	.0000	.0000	.0000
.0000	.0000	.0000	.0000	.0000	.0000
.0000	.0000	.0000	.0000	.0000	.0000
.0000	.0000	.0000	.0000	.0000	.0000
.0000	.0000	.0000	.0000	.0000	.9000+02

c) With Constraints (means)

NO.OF MISCLASSIFICATIONS IN CLASS 1=			6		
NO.OF MISCLASSIFICATIONS IN CLASS 2=			5		
NO.OF MISCLASSIFICATIONS IN CLASS 3=			0		
NO.OF MISCLASSIFICATIONS IN CLASS 4=			2		
NO.OF MISCLASSIFICATIONS IN CLASS 5=			2		
NO.OF MISCLASSIFICATIONS IN CLASS 6=			5		
-.1000+02	.0000	.0000	.0000	.0000	.0000
.0000	-.8000+01	.0000	.0000	.0000	.0000
.0000	.0000	-.3000+02	.0000	.0000	.0000
.0000	.0000	.0000	-.2000+02	.0000	.0000
.0000	.0000	.0000	.0000	-.4000+02	.0000
.0000	.0000	.0000	.0000	.0000	-.1000+02

Thus, the percentage of misclassification was reduced from 33.95% to 9.5%. Another possibility is a parametric search for the matrix-- $M^T\hat{W} - F$ which involves mainly the diagonal elements. Other constraints are possible by utilizing the means of the misclassified samples or the samples that are misclassified and have the largest mean-square-error.

CHAPTER V

THE WEIGHTED MEANS

We suggest here utilizing the means of the classes using the weighting approach suggested in Chapter V. We noted in Chapter IV that the MSE results in weight vectors proportional to the means of the classes and that a broad class of discriminant functions yield weight vectors proportional to the means of the classes. This procedure is easily programmed and computationally economical.

Let A consist of the sample means of each class:

$$A = \begin{bmatrix} m_1^T \\ \vdots \\ m_K^T \end{bmatrix} \quad , \qquad B = \begin{bmatrix} 0 & \gamma & \cdots & \gamma \\ \gamma & 0 & \cdots & \gamma \\ \gamma & \gamma & \cdots & \vdots \\ \vdots & \vdots & \cdots & \gamma \\ \gamma & \gamma & \cdots & 0 \end{bmatrix}$$

$$\therefore \qquad A^T B = [m_1 \ \cdots \ m_K] \begin{bmatrix} 0 & \gamma & \cdots & \gamma \\ \gamma & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma \\ \gamma & \cdots & \gamma & 0 \end{bmatrix}$$

$$A^T B = \left[\gamma \sum_{i \neq 1} m_i \qquad \gamma \sum_{i \neq 2} m_i \quad \cdots \quad \gamma \sum_{i \neq K} m_i \right]$$

$$= \gamma[-m_1 + m \qquad -m_2 + m \quad \cdots \quad -m_K + m]$$

and

$$\hat{W}^{(i)} = -(A^T A)^{-1} m_i$$

where

$$A^T A = \sum_{i=1}^{k} m_i \, m_i^T \quad .$$

The procedure of weighting used in Chapter II is utilized here by repeating the means of the classes that have the most misclassified samples. The procedure is terminated either after specified number of iterations or after a satisfactory result is obtained.

The procedure was carried on the data base of Sebestyen and Edie [47] that we used earlier and the results are shown below:

2 features (x_1, x_2)

Class	No. Misc.	Weighting	MSE No. Misc.
1	0	|3, 2, 2, 2, 2, 3|	6
2	2		0
3	0		0
4	0		0
5	3		18
6	25		33
	30		57

3 features (x_1, x_2, x_1x_2)

Class	No. Misc.	Weighting	MSE No. Misc.
1	1	|3, 5, 2, 2, 2, 7|	6
2	5		0
3	1		0
4	2		2
5	2		2
6	12		30
	23		40

Where the first entry in the weighting is the number of times the mean of the first class was repeated, the second entry is the number of times the mean of the second class was repeated, etc.

Limitations

For the weighting method to work, the set of equations have to be inconsistent (Theorem 1, Chapter II). In our case we have six classes, hence the matrix A is $m \times n$ with $m = 6$. If $n \geq 6$ with the rank of $A = 6 < n$, then the equations are consistent. If $n = 6$ and the rank of $A = 6$, then we have a unique solution, the inverse of A exists and the equations are consistent. Hence, we expect the weighting method will not work with $n = 6$; i.e., 5 unaugmented features. Actually we had an almost consistent set of equations with 4 features (5 augmented).

The Means and the Means of the Errors

Because of the limitations stated above we resorted to introducing the means of the patterns that are in error in each class in addition to the means of the patterns of each class. The same weighting method is applied now where the repetition would be repetition of the means and the errors of the means depending on whether the misclassified samples lie in those that were in error or not. The results are shown below:

2 features (x_1, x_2)

Class	No. Misc.	Weighting	MSE No. Misc.
1	0	|3,0,2,0,2,0,2,0,2,0,1,2|	6
2	1		0
3	0		0
4	0		0
5	3		18
6	18		33
	22		57

3 features (x_1, x_2, x_1x_2)

Class	No. Misc.	Weighting	MSE No. Misc.
1	2	|5,0,3,2,2,2,2,1,2,1,8,3|	6
2	6		0
3	0		0
4	2		2
5	2		2
6	6		30
	18		40

Where the first entry in the weighting is the number of times the mean of the first class was repeated, the second entry is the number of times the mean of the samples in the first class that were misclassified by weighting the means only was repeated. The third entry is the number of times the mean of the second class was repeated, and the fourth entry is the number of times the mean of the samples in the second class that were misclassified by weighting the mean only was repeated, etc. The results are summarized in Table 5.

This procedure can be considered as a special case of MacQueen K-means clustering approach [34] and Sebestyen's adaptive sample set construction [55], which are clustering methods aiming at representing a large number of samples by smaller number representative of their classes. The representative samples selected must yield decision boundaries very much like those obtained by the larger number of the given samples. This suggests a problem-oriented approach for clustering utilizing the means of the samples correctly classified and the means of the samples that are in error. This will not be pursued further here.

Table 5. Comparison of the Weighted Means and MSE Results.

	No. of Misclassification	% of Misclassification
MSE Solution 2 features	57	33.95
Weighted Means 2 features	30	17.85
Weighted Means and the Means of the Errors 2 features	22	13.1
MSE Solution 3 features	40	23.8
Weighted Means 3 features	23	13.7
Weighted Means and the Means of the Errors 3 features	18	10.7

CHAPTER VI

APPLICATION OF GENERALIZED INVERSE TO FEATURE EXTRACTION

In feature extraction there are two points of view --One saying that the purpose of feature extraction is to reduce the number of measurements and hence the cost of the measurements; the other, while conceding the above point, says that even after deciding on the number of measurements to be taken it is essential to reduce these by processing them into more discriminating features using different transforms. Many of these transforms are Fourier expansion or Karhunen-Loéve expansion types. The latter point of view gains importance with Foley's recent results [15]. Foley shows that if m arbitrary samples are randomly thrown down according to uniform distribution in n-dimensional space (n-features), then a certain optimal linear classification seems to indicate that the categories are widely separated unless $\frac{m}{n}$ is approximately three or more. Since the patterns from both sets are drawn according to the same distribution, this could lead the experimenter to spurious conclusion. In the following pages we relate Fourier expansion, Karhunen-Loéve expansion, and the generalized inverse.

A natural relationship exists between the generalized inverse, the Fourier expansion, and the Karhunen-Loéve expansion. This relation stems from the fact that all three techniques involve least-square approximations by projection onto a subspace.

The Moore-Penrose least-square generalized inverse (pseudo-inverse)

is well defined for bounded linear operators with closed range. Let H_1 and H_2 be Hilbert spaces and $L(H_1,H_2)$ the set of bounded linear transformation from H_1 into H_2 with closed range. For each A in $L(H_1,H_2)$ the following orthogonal decomposition is well known [32].

$$H_1 = R(A^*) + N(A) \quad , \quad H_2 = R(A) + N(A^*) \tag{1}$$

Where R(A) denotes the range of A, N(A) denotes the null space of A, A^* denotes the adjoint of A, and + denotes the direct sum. Each element u in H_2 can be expressed as $u = u_1 + u_o$ where $u_1 \in R(A)$ and $u_o \in N(A^*)$. Similarly each element x in H_1 can be expressed as $x = x_1 + x_o$ where $x_1 \in R(A^*)$ and $x_o \in N(A)$. Let $P_{R(A)}$ be the transformation defined by $P_{R(A)}: u \to u_1$. It follows that $P_{R(A)}$ is an orthogonal projector onto R(A). $P_{N(A^*)}$, $P_{R(A^*)}$, and $P_{N(A)}$ are similarly defined and can be expressed in terms of the pseudo inverse $A^\dagger$ as follows [36]

$$\begin{aligned} P_{R(A)} &= AA^\dagger \quad , \quad P_{N(A^*)} = I - AA^\dagger \\ P_{R(A^*)} &= A^\dagger A \quad , \quad P_{N(A)} = I - A^\dagger A \end{aligned} \tag{2}$$

where the pseudo-inverse $A^\dagger$ is defined as the unique solution X of the following equations [46]

$$XAX = X \tag{3}$$

$$(AX)^* = AX \tag{4}$$

$$AXA = A \tag{5}$$

$$(XA)^* = XA \tag{6}$$

and has the properties:

$$||AA^{\dagger}u - u|| \leq ||Ax - u|| \quad \text{for all } x \in H_1$$

and (7)

$$||A^{\dagger}u|| \leq ||x|| \quad \text{if} \quad ||Ax - u|| = ||AA^{\dagger}u - u||$$

where $||\cdot||$ denotes the Euclidean norm. In other words, a least means square solution of $Ax = u$ is $\hat{x}_1 = A^{\dagger}u$ and this solution is the unique solution of minimum norm. Substituting $\hat{x}_1 = A^{\dagger}u$ for x in the minimized expression $\min_x ||Ax - u|| = ||AA^{\dagger}u - u|| = ||P_{R(A)}\, u - u||$.

In the Fourier expansion we seek the least-square best approximation for

$$||u - \sum_{i=1}^{k} \alpha_i \varphi_i|| \tag{8}$$

where u is a vector in a space of dimension $\geq k$. φ_i-s are a set of orthonormal basis in k-dimensional space. The minimum is obtained by choosing $\alpha_i = \langle u, \varphi_i \rangle$, where $\langle \cdot , \cdot \rangle$ denotes the inner product.

Define

$$Ax = \sum_{i=1}^{k} \alpha_i \varphi_i \quad , \qquad x = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix} \tag{9}$$

or

$$Ax = \langle x, \Phi \rangle = x^T \Phi \quad \text{where} \quad \Phi = \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_k \end{pmatrix}$$

So the problem is to minimize: $||u - Ax||$ over $x \in R^k$. (10)

The general solution to (10) is:

$$\hat{x} = A^{\dagger}u + (I - A^{\dagger}A)x \tag{11}$$

where x is an arbitrary element in R^k and $(I - A^\dagger A) = P_{N(A)}$. Since $\{\varphi_i\}$ are linearly independent, N(A) = 0. So in this case a unique solution is obtained as

$$\hat{x} = A^\dagger u \qquad \text{and} \qquad AA^\dagger u = \sum_{i=1}^{k} \langle u, \varphi_i \rangle \varphi_i = P_{R(A)} u \tag{12}$$

In Karhunen-Loéve expansion [37] we seek the least-square best approximation for

$$\left|\left| u(t) - \sum_{i=1}^{k} b_i \varphi_i(t) \right|\right| \tag{13}$$

where u(t) is a random stochastic process over the time period (o,T). The coefficients b_i are orthogonal. Assuming $E\{u(t)\} = 0$, then $E\{b_i\} = 0$. In this case orthogonality of the random variable b_i is equivalent to uncorrelatedness so that

$$\begin{aligned} E\{b_i b_j^*\} &= 0 \qquad i \neq j \\ E\{b_i^2\} &= \lambda_i \end{aligned} \tag{14}$$

If the process u(t) is stationary we have

$$\int_o^T R(t - \tau) \varphi_i(\tau) d\tau = \lambda_i \varphi_i(t) \tag{15}$$

where $R(t, t) = E\{u^2(t)\}$.

$\{\varphi_i(t)\}$ is a set of deterministic orthonormal coordinates functions over (o, T), i.e.,

$$\int_o^T \varphi_i(t) \varphi_j^*(t) dt = \delta_{ij} \tag{16}$$

where δ_{ij} is the Kronecker delta-function, equal to one if i = j and zero otherwise.

Define

$$Ax = \sum_{i=1}^{k} b_i \varphi_i(t) \quad , \qquad x = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} \tag{17}$$

$$b_i = \int_0^T u(t)\varphi_i^*(t)dt = \langle u(t), \varphi_i(t) \rangle \ , \ t\varepsilon(o,T)$$

So the problem is to minimize $||u(t) - Ax||$ over x. The generalized inverse solution is

$$\hat{x} = A^{\dagger} u(t) \tag{18}$$

therefore,

$$A\hat{x} = AA^{\dagger}u(t) = P_{R(A)}u(t)$$

$$= \sum_{i=1}^{k} \langle u(t), \varphi_i(t) \rangle \varphi_i(t) \tag{19}$$

since R(A) has $\varphi_i(t)$, i =1, ..., k for basis.

Thus, the relations between the generalized inverse approach, Fourier series expansion, and Karhunen-Loéve expansion are shown.

In view of the prevalent use of the Fourier expansion and Karhunen-Loéve expansion in feature extraction, it is the author's contention that a closer look should be taken at the possibility of applying generalized inverse techniques in feature extraction.

We take the point of view that the effectiveness of features could only be measured by the particular classifier that we are going

to use. Our results using weighting and constraints compare very favorable with Wee's results in feature extraction on the same data base of Sebestyen and Edie [55]. Up to eighth order polynomials were used to constitute the feature set to be selected. The results is shown in Fig. 5 with curve A representing the performance using the selection procedure, and curve B representing the performance when the features were taken in natural sequence as x_1, x_2, x_1^2, x_1x_2, x_2^2, etc. Five features [x_2, $x_1^3 x_2$, x_1^3, x_2^3, $x_1x_2^2$] were needed to give 90 percent recognition. We got over 90 percent recognition using just two features, x_1 and x_2.

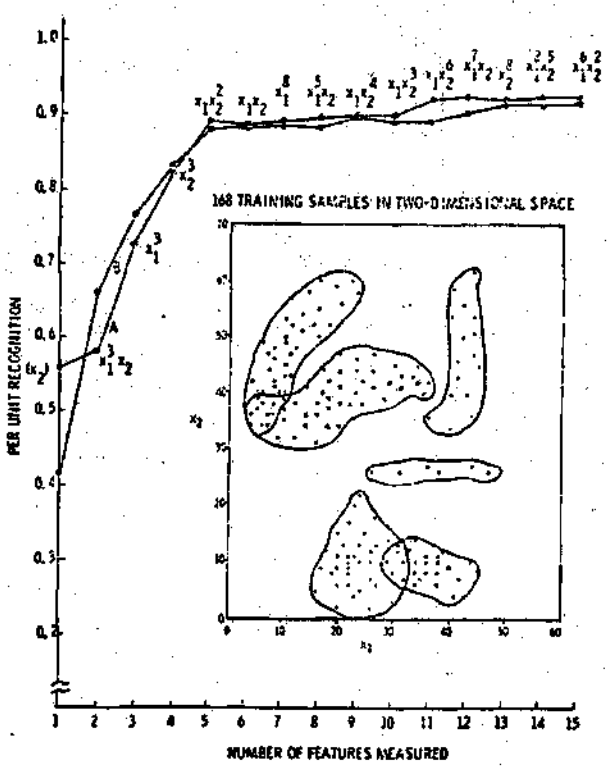


Figure 5. Wee's Feature Extraction Results.

CHAPTER VII

SOME COMPUTATIONAL ASPECTS

In the computations carried out on the data base by Sebestyen and Edie [55], the resulting matrix A is of full column rank and hence $A^{\dagger} = (A^{T}A)^{-1}A^{T}$. The Gauss-Jordan elimination method available in the math pack of the 1108 Univac was utilized with no difficulties encountered. The CPU time per program was about five CPU seconds for all computations. In this chapter some of the reservations about the computation of $(A^{T}A)^{-1}$ is going to be discussed. Kishi's algorithm which is particularly suitable to the iterative scheme suggested in this dissertation is going to be discussed in more detail.

The main objection to direct computation of $A^{\dagger}$ from $A^{\dagger} = (A^{T}A)^{-1}A^{T}$ is that if the original problem is ill-conditioned, the method will make the condition worse [43]. The condition number for A is given by

$$K(A) = ||A||_2 \, ||A^{\dagger}||_2$$

$$K(A^{T}A) = [K(A)]^2 \quad .$$

The linear equation $Aw = b$, and the matrix A, are said to be ill-conditioned if the solutions are very sensitive to small changes in the data. The following example illustrates the ill-conditioning of the normal equation.

Let

$$A = \begin{bmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix}$$

and let the elements of

$$A^T A = \begin{bmatrix} 1 + \varepsilon^2 & 1 \\ 1 & 1 + \varepsilon^2 \end{bmatrix}$$

be computed using double-precision and then rounded to single precision using t binary digits. If $|\varepsilon| < \sqrt{2^{-t}}$ then the

$$fl(A^T A) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad \text{(fl denotes floating point)}$$

which is of rank 1, whereas A is of rank 2. Thus, the computed normal equation

$$fl(A^T A)w = fl(A^T b)$$

may have solutions which are not least square solutions of $Aw = b$. Suggested methods to avoid the worsening of the condition are LU decomposition and orthogonal decomposition of A. For more detailed discussion on this point see Noble [42], [43] and chapter 3 of Ben Israel and Greville [6].

There are many algorithms for the computation of the generalized inverse and the least square solutions. The handbook for automatic computation by Wilkinson and Reinsch [72] provides a summary and algol computer programs for many of these. Shinozaki et al. provide a critique for many of these algorithms in their two papers [56], [57]. In the remaining part of this chapter Kishi's algorithm, which is particularly suitable to our iterative methods, will be discussed.

Kishi's algorithm [29] is the dual of Greville's algorithm [20]. Greville's algorithm is a recursive algorithm for finding the solution of minimum norm $\hat{w} = A^{\dagger}b$ that minimizes the Euclidean norm of the error $e = ||Aw - b||$ by recursively adding the columns of A while Kishi's algorithm recursively adds the rows of A.

Kishi showed the quivalence between his algorithm and Kalman's procedure [26], [27] applied to the time independent case. Ben Israel and Greville [6] noted the equivalence between Albert and Stitler derivation of Kalman's for the time independent case and Kishi's algorithm [3].

Kishi's Algorithm

Let

$$A_K w_K = b_K$$

At each occurrence of data acquisition a new component of data is appended to the vector b_K. For this case $\{w_K\}$, $(K = 1,2,\ldots)$, is a sequence of vectors such that each vector w_K is the solution of minimum norm based upon K pieces of data.

Kishi's iterative method is of the form

Let

$$A_{K+1} = \begin{pmatrix} A_K \\ * \\ a_{K+1} \end{pmatrix} \quad , \quad b_{K+1} = \begin{pmatrix} b_K \\ b_{K+1} \end{pmatrix}$$

$$\hat{w}_{K+1} = A^{\dagger}_{K+1}\, b_{K+1} = (B_{K+1} \, , \, h_{K+1}) b_{K+1} \, .$$

The Procedure

Let

$$c^{*}_{K+1} = a^{*}_{K+1} - a^{*}_{K+1}\, A^{\dagger}_{K}\, A_K \quad .$$

Case 1: $c_{K+1} \neq 0$

$$h_{K+1} = c_{K+1} \, (c^*_{K+1} \, c_{K+1})^{-1}$$

Case 2: $c_{K+1} = 0$

$$h_{K+1} = (1 + a^*_{K+1} \, A^\dagger_K \, A^{\dagger *}_K \, a_{K+1})^{-1} \, A^\dagger_K \, A^{\dagger *}_K \, a_{K+1} \quad .$$

Then

$$\hat{w}_{K+1} = \hat{w}_K - h_{K+1} \, a^*_{K+1} \, \hat{w}_K + h_{K+1} \, b_{K+1} \quad .$$

Also,

$$A^\dagger_{K+1} \, A_{K+1} = A^\dagger_K \, A_K + h_{K+1} \, c^*_{K+1}$$

$$A^\dagger_{K+1} \, A^{\dagger *}_{K+1} = (b_{K+1} \, a^*_{K+1} - I) \, A^\dagger_K \, A^{\dagger *}_K (I - (h_{K+1} \, a^*_K)^*) + h_{K+1} \, h^*_{K+1} \quad .$$

The flow chart for the compution is shown in Fig. 6.

Kishi's algorithm gives a method of computing the best estimate in terms of the previously calculated best estimate. This computation is performed in a manner which saves computer storage and in a manner not requiring matrix inversion. The procedure will always give a solution since Penrose have shown the existence and uniqueness of the generalized inverse.

In our case the initial solution could be achieved using any method and Kishi's procedure could be used for the proceeding iterations.

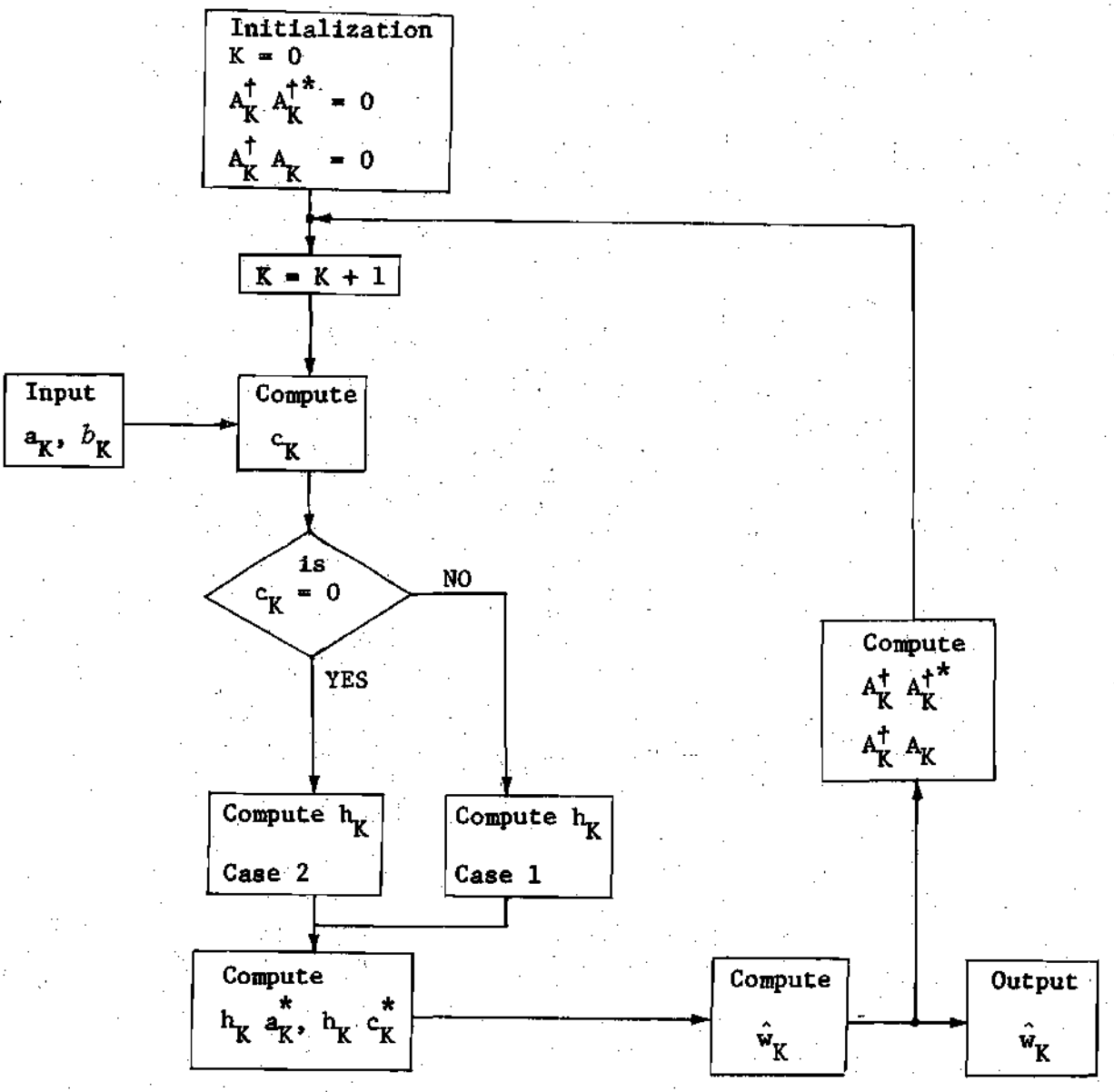


Figure 6. Flow Chart for Recursive Method.

CHAPTER VIII

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

Several results were presented in this dissertation applying the generalized inverse concepts and techniques to pattern recognition and introducing new results. A new weighted MSE approach to pattern classification was introduced and motivated statistically. Experimental results on nonseparable classes were very favorable. The approach yielded a new algorithm for pattern classification that converges to a separating solution if the patterns are linearly separable. The new algorithm is remarkable in the sense that it is one of the few descent algorithms derived from pattern recognition considerations. Yet, it bears a great deal of similariities to other algorithms. The other algorithms are usually brought in from the literature on inequalities. An adaptive constraint procedure for pattern classification was presented with very favorable experimental results. Also, the weighting approach was applied to the means of the classes and the means of the samples that were misclassified. The experimental results on the weighted means were equally favorable suggesting a problem oriented clustering technique. Finally, it was noted that the experimental results of this dissertation were better than the experimental results for feature extraction on the same data base and concluded that the techniques of this dissertation could be used for feature extraction

purposes. This was followed up by pointing out the relation between the generalized inverse, Fourier and Karhunen-Loéve expansions. The latter expansions are used often for feature extraction purposes.

Recommendations

An aspect of the new algorithm that could be pursued further is its rate of convergence. While many of the existing algorithms lack a study of their rate of convergence, the new algorithm rate of convergence could be more tractable.

In view of the experimental results and the relation between the generalized inverse and other transforms used in feature extraction, further investigations on applying the techniques presented in this dissertation to feature extraction is thought worth pursuing.

In the computations undertaken in this dissertation no difficulties were encountered in the computation of the inverse of (A^TA) resulting from possible worsening of the condition number. The adverse effects of the worsening of the condition number results when a small perturbation of the matrix A causes a drastic change in the solution. This does not occur in pattern recognition since a small perturbation of the matrix A corresponds to a small perturbation of the given samples and this should not change the classification drastically. Further study of this point would be worthwhile.

While the relation between the MSE pattern classifier and the expected loss is one of the best well-known unsolved problems, it is hoped that the results of this dissertation could serve as a step in the direction of the solution of that problem.

APPENDICES

APPENDIX A

THE GENERALIZED INVERSE (THE MOORE-PENROSE INVERSE)

In 1920, Moore [37] introduced the notion of a generalized inverse for singular or rectangular matrices. Moore's various results were later incorporated in [38]. These results were not well known and generalized inverses were defined later independently by Penrose [46], Bjerhammer [7] and others. Greville [19], [20] gave an impetus to the study of generalized inverses. Later, the concept of generalized inverses of linear operators in the setting of functional analysis has emerged [40].

At the present time, the theory is elegant, the applications are diverse (e.g., least-squares, linear equations, projections, statistical regression analysis, filtering and linear programming) and most important, a deeper understanding of these topics is achieved when they are studied in the generalized inverse context (Albert [2]; Rao and Mitra [52]; Boullion and Odell [9]; Pringle and Rayner [51]; Bjerhammer [8]).

Example

Let us consider the system of linear equations

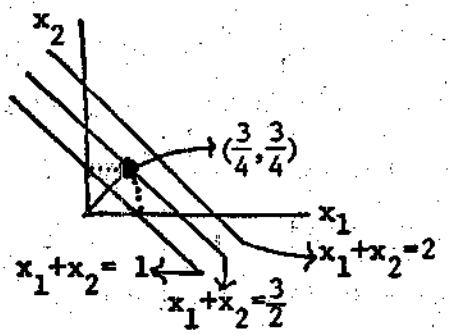
$$Ax = y, \quad \text{where } A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \; ; \; y = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{1}$$

This system obviously has no solution in the traditional sense; the

vector y is not in the range of the matrix A.

One may broaden the notion of a solution so that (1), and indeed any system of linear algebraic equations has a solution in a meaningful sense. One way to do this is to replace the vector y in (1) by another vector $\tilde{y}$ which satisfies two requirements: $\tilde{y} \in R(A)$, the range of A, and $\tilde{y}$ approximate y in some sense; for instance we may take y to be the point in R(A) "nearest" to y. If the usual Euclidean distance is used, then $\tilde{y}$ is obviously the orthogonal projection of y in R(A). Then the equation

$$Ax = P_{R(A)}\, y = \tilde{y} \tag{2}$$



where $P_{R(A)}$ denotes the orthogonal projection on R(A) is solvable, but the solution is not unique. Note that R(A) = Span $\{(1,1)^T\}$, $P_{R(A)}y = (\frac{3}{2}, \frac{3}{2})^T$, and thus the set of all solutions of (2) is given by $S = \{(x_1, x_2): x_1 + x_2 = \frac{3}{2}\}$.

Another way of attaching a notion of solution to (1) is to seek a solution in the least-squares sense. A vector $u \in R_2$ is called a least-squares solution of (1) if

$$\min_x \{||Ax - y||: x \in R_2\} = ||Au - y|| \quad , \tag{3}$$

where $||\cdot||$ is some norm on R_2. It is easy to show that corresponding

to the choice of the Euclidean norm $||x|| = (x_1^2 + x_2^2)^{\frac{1}{2}}$, a vector $u \in R_2$ is a solution of (2) if and only if u is a least-squares solution of (1). The problem of minimizing $||Ax - y||^2$ in the Euclidean norm is also equivalent to solving the equation

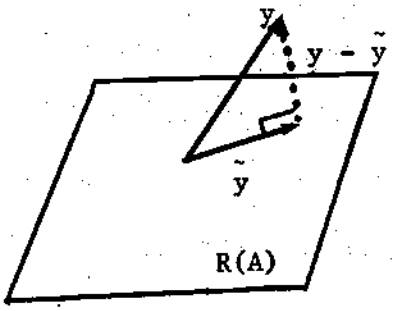
$$A^*Ax = A^*y \quad , \tag{4}$$

where A* is the adjoint of A (transpose in real finite dimensional spaces). (Equation (4) is often called the "normal" equation by analogy with the normal equation arising in least-squares problems in statistics).

If we seek a vector of minimal Euclidean norm which minimizes $||Ax - y||$, we get the unique solution $\hat{x} = (\frac{3}{4}, \frac{3}{4})$. This analysis leads to an analytic definition of the Moore-Penrose generalized inverse of A. More explicitly, $A^\dagger$ is the map on R_2 into R_2 which assigns to each $y \in R_2$, the vector $\hat{x}$ which minimizes $||Ax - y||$ and has the property that $||\hat{x}|| < ||u||$ for all u with $||A\hat{x} - y|| = ||Au - y||$. For the example considered, we thus have $R(A) = \text{span}\begin{pmatrix}1\\1\end{pmatrix}$; let

$$v = \frac{\begin{pmatrix}1\\1\end{pmatrix}}{\left|\left|\begin{pmatrix}1\\1\end{pmatrix}\right|\right|} = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix} \qquad \tilde{y} = P_{R(A)}\, y = \langle y, v\rangle\, v = \begin{pmatrix}\frac{3}{2}\\[2pt]\frac{3}{2}\end{pmatrix}$$

$$A^\dagger = \begin{bmatrix}\frac{1}{4} & \frac{1}{4}\\[2pt]\frac{1}{4} & \frac{1}{4}\end{bmatrix} \quad \text{and} \quad \hat{x} = A^\dagger y = \begin{bmatrix}\frac{3}{4}\\[2pt]\frac{3}{4}\end{bmatrix}$$



where $\langle y, v\rangle = y^*v$ is the inner product.

Definition

Among all vectors x_1 satisfying

$$||Ax_1 - y|| = \min_x ||Ax - y|| \quad ; \quad x \in R_n$$

let $\hat{x}$ be the unique vector of minimum norm. The generalized inverse $A^\dagger$ of A is the matrix mapping y into its corresponding $\hat{x}$.

The general least-square solution is

$$x_1 = A^\dagger y + z \quad , \qquad z \in N(A)$$

where z is in the null space of A.

$$Ax_1 - y = AA^\dagger y - y = \tilde{y} - y$$

Thus $\hat{x} = A^\dagger y$ and $AA^\dagger = P_{R(A)}$.

The generalized inverse has the following important properties:

1) It always exists for finite dimensional spaces, since the range is closed and the projection on it exists.
2) It is unique.
3) It is linear.

The generalized inverse was characterized by Penrose as the unique solution X of the following four matrix equations [46]

(i) $XAX = X$ (ii) $AXA = A$

(iii) $(AX)^* = AX$ (iv) $(XA)^* = XA$

Further properties of the generalized inverse:

a) $(A^\dagger)^\dagger = A$

b) $(A^*)^\dagger = (A^\dagger)^*$

c) $A^\dagger = (A^*A)^\dagger A^*$

d) $A^{\dagger} = A^{*} (AA^{*})^{\dagger}$

e) If A is nonsingular $A^{\dagger} = A^{-1}$

f) If A is m x n, then $A^{\dagger}$ is n x m and has its rows and columns in the row-space and column-space of A^{*}

g) If A is of full column-rank, then $A^{\dagger} = (A^{*}A)^{-1} A^{*}$

h) If A is of full row-rank, then $A^{\dagger} = A^{*} (AA^{*})^{-1}$

i) If A is of rank $r \leq \min(m,n)$ we can express A as a product A = BC, where B is m x r and C is r x n, and both of rank r, then

$$A^{\dagger} = C^{*} (CC^{*})^{-1} (B^{*}B)^{-1} B^{*} = C^{\dagger}B^{\dagger}$$

For $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, choose $B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $C = (1 \quad 1)$, then

$$A^{\dagger} = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

j) $(AB)^{\dagger} = B^{\dagger}A^{\dagger}$ if and only if both the equations,

(1) $A^{\dagger}AB\,(AB)^{*} = B(AB)^{*}$

(2) $B\,B^{\dagger}A^{*}AB = A^{*}AB$

are satisfied.

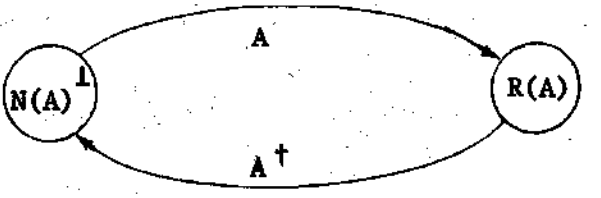
k) $(A^{*}A)^{\dagger} = A^{\dagger}A^{\dagger *}$

l) A, $A^{*}A$, $A^{\dagger}$ and $A^{\dagger}A$ all have rank equal to trace $A^{\dagger}A$.

We also have the following orthogonal projectors,

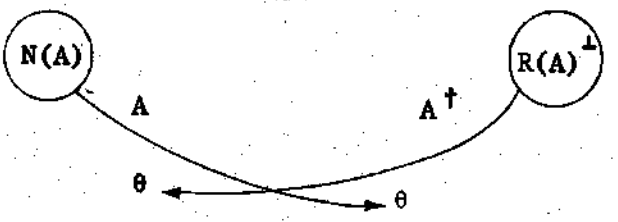
$P_{R(A)} = AA^{\dagger}$

$P_{R(A^{*})} = A^{\dagger}A$



$$P_{N(A^*)} = I - AA^\dagger$$

$$P_{N(A)} = I - A^\dagger A$$



APPENDIX B

SEBESTYEN AND EDIE'S DATA

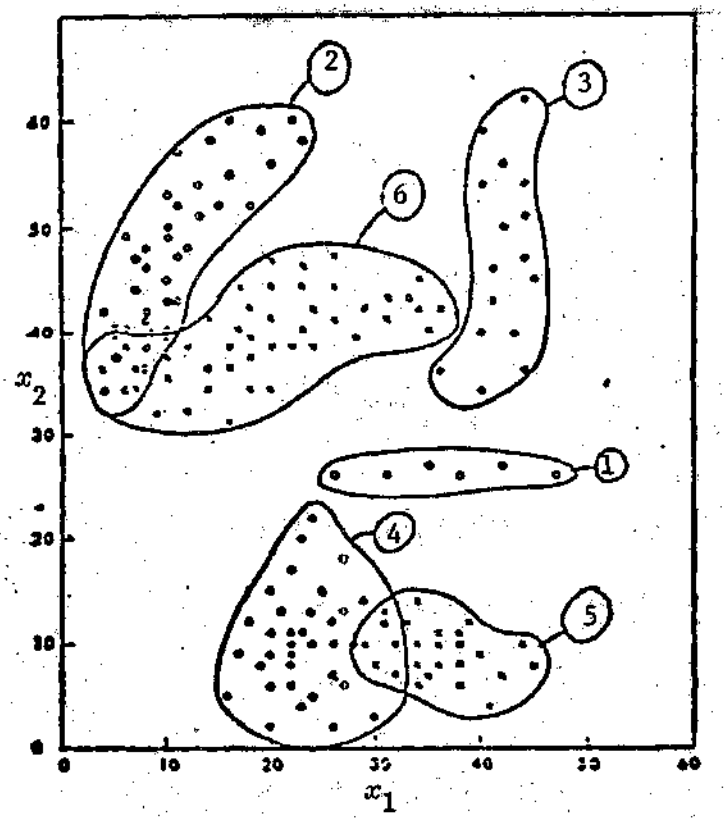


Figure 3. 168 Two-Dimensional Vectors.

In this appendix our reading of Sebestyen and Edie's data, Fig. 3, is presented for easy reference and verification of our results by the the reader. Sebestyen and Edie [54] claimed 170 two-dimensional vectors. Wee [68] could count only 168 samples. We, too, could count only 168 vectors. All the samples have integer values and our reading of them agrees with Wee's reading since we were able to reproduce his results. The sample values for the six classes are presented below.

Note that in our computations each vector was augmented by adding a third coordinate of values +1 to it.

		x_1	x_2
Class 1 (6 Samples)	1.	47	27
	2.	42	28
	3.	38	27
	4.	35	28
	5.	31	27
	6.	28	27
Class 2 (39 Samples)	7.	4	42
	8.	4	34
	9.	5	40
	10.	5	37
	11.	6	49
	12.	6	40
	13.	6	34
	14.	7	47
	15.	7	44
	16.	8	48
	17.	8	46
	18.	8	42
	19.	8	41
	20.	8	38
	21.	8	36
	22.	9	40
	23.	10	53
	24.	10	50
	25.	10	49
	26	10	45
	27	10	43
	28	10	40
	29	11	57
	30	11	52
	31	11	47
	32.	11	43
	33.	12	48
	34.	13	54
	35.	13	51
	36.	14	58
	37.	15	52
	38.	16	60
	39.	16	55
	40.	16	50
	41.	18	52
	42.	19	59
	43.	20	56
	44.	22	60
	45.	24	58

		x_1	x_2
Class 3 (16 Samples)	46.	36	36
	47.	40	34
	48.	40	40
	49.	40	54
	50.	40	59
	51.	41	43
	52.	41	46
	53.	42	50
	54.	42	56
	55.	43	40
	56.	44	36
	57.	44	47
	58.	44	51
	59.	44	54
	60.	44	62
	61.	45	45
Class 4 (37 Samples)	62.	16	5
	63.	17	9
	64.	18	12
	65.	19	8
	66.	20	15
	67.	20	11
	68.	20	9
	69.	20	6
	70.	20	2
	71.	21	13
	72.	22	17
	73.	22	11
	74.	22	10
	75.	22	9
	76.	22	8
	77.	22	6
	78.	23	20
	79.	23	11
	80.	23	4
	81.	23	0
	82.	24	22
	83.	24	13
	84.	24	10
	85.	24	5
	86.	25	15
	87.	26	12
	88.	26	10
	89.	26	7
	90.	26	2

		x_1	x_2
Class 4 (Continued)	91.	27	18
	92.	27	13
	93.	27	6
	94.	28	10
	95.	29	14
	96.	30	3
	97.	31	12
	98.	32	7
Class 5 (23 Samples)	99.	29	10
	100.	30	8
	101.	31	13
	102.	32	10
	103.	33	12
	104.	34	14
	105.	34	10
	106.	34	8
	107.	34	6
	108.	35	7
	109.	36	11
	110.	36	10
	111.	36	8
	112.	38	11
	113.	38	10
	114.	38	8
	115.	38	6
	116.	39	12
	117.	40	9
	118.	41	4
	119.	42	7
	120.	44	10
	121.	45	8
Class 6 (47 Samples)	122.	4	36
	123.	5	39
	124.	6	38
	125.	7	36
	126.	7	34
	127.	9	32
	128.	10	39
	129.	10	38
	130.	10	36
	131.	12	38
	132.	12	32
	133.	14	41
Class 6 (Continued)	134.	14	36
	135.	14	34
	136.	16	38
	137.	16	36
	138.	16	31
	139.	17	44
	140.	17	40
	141.	18	42
	142.	18	40
	143.	18	38
	144.	18	34
	145.	20	46
	146.	20	44
	147.	20	43
	148.	20	38
	149.	20	34
	150.	22	38
	151.	23	46
	152.	23	44
	153.	23	40
	154.	24	42
	155.	24	38
	156.	26	47
	157.	26	44
	158.	26	41
	159.	28	39
	160.	29	42
	161.	30	45
	162.	31	43
	163.	31	41
	164.	33	43
	165.	34	45
	166.	34	42
	167.	35	40
	168.	36	42

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Agmon, S., "The Relaxation Method for Linear Inequalities," Canadian Journal of Mathematics, Vol. 6, 1954, pp. 382-392.

2. Albert, A., Regression and the Moore-Penrose Pseudo-Inverse, Academic Press, New York, 1972.

3. Albert, A. and K. W. Stittler, "A Method for Computing Least Squares Estimators that Keep Up with the Data," SIAM Journal on Control, Vol. 3, No. 3, 1966, pp. 384-417.

4. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1958.

5. Andrews, H. C., Introduction to Mathematical Techniques in Pattern Recognition, Wiley, New York, 1972.

6. Ben-Israel, A. and T. N. E. Greville, Generalized Inverses: Theory and Applications, Wiley, New York, 1974.

7. Bjerhammer, A., "Rectangular Reciprocal Matrices with Special Reference to Geodetic Calculations," Bulletin Geodesique, 1951, pp. 188-220.

8. Bjerhammer, A., Theory of Errors and Generalized Matrix Inverses, Elsevier Scientific Publishing Company, Amsterdam, Holland, 1973.

9. Boullion, T. L. and P. L. Odell, Generalized Inverse Matrices, Wiley, New York, 1971.

10. Businger, P. and G. H. Golub, "Linear Least Squares Solution by Householder Transformation," Numer. Math. 7, 1965, pp. 269-276.

11. Chien, Y. T. and K. S. Fu, "Selection and Ordering of Feature Observations in a Pattern Recognition System," Information and Control, Vol. 12, May-June 1968, pp. 394-414.

12. Dertouzos, M. L., "An Approach to Single Threshold Elements Synthesis," IEEE Trans. Computers, Vol. EC-13, October 1964, pp. 519-528.

13. Duda, R. O. and P. E. Hart, Pattern Classification and Scene Analysis, John Wiley, New York, 1973.

BIBLIOGRAPHY (Continued)

14. Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," Ann. Eugenics, 7, Part II, 1936, pp. 179-188; also in Contributions to Mathematical Statistics, John Wiley, New York, 1950.

15. Foley, D. H., "Considerations of Sample and Feature Size," IEEE Trans. Info. Theory, Vol. IT-18, No. 5, September 1972, pp. 618-626.

16. Fu, K. S., Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, New York, 1968.

17. Fukunaga, K., Introduction to Statistical Pattern Recognition, Academic Press, New York, 1972.

18. Golub, G. H. and C. Reinsch, "Singular Value Decomposition and Least Squares Solutions," Numer. Math. 14, 1970, pp. 403-420.

19. Greville, T. N., "The Pseudo Inverse of a Rectangular or Singular Matrix and Its Application to the Solution of Systems of Linear Equations," SIAM Review 1, 1959, pp. 38-43.

20. Greville, T. N., "Some Applications of the Pseudo Inverse of a Matrix," SIAM Review 2, 1960, pp. 15-22.

21. Hanson, R. J. and C. L. Lawson, "Extensions and Applications of the Householder Algorithm for Solving Linear Least Squares Problems," Mathematics of Computation, October 1969, Vol. 23, No. 108, pp. 787-812.

22. Highleyman, W. H., "Linear Decision Functions with Application to Pattern Recognition," Proc. IRE, 50, June 1962, pp. 1501-1514.

23. Ho, Y. C. and A. K. Agrawala, "On Pattern Classification Algorithms Introduction and Survey, Proceedings of the IEEE, Vol. 56, No. 12, December 1968, pp. 2101-2114.

24. Ho, Y. C. and R. L. Kashyap, "A Class of Iterative Procedures for Linear Inequalities," J. SIAM Control, 4, 1966, pp. 112-115.

25. Ho, Y. C. and R. L. Kashyap, "An Algorithm for Linear Inequalities and Its Applications," IEEE Trans. Elec. Comp., EC-14, October 1965, pp. 683-688.

26. Kalman, R. E., "New Results in Linear Filtering and Prediction Theory," Trans. ASME, Series D, Journal of Basic Engineering, 1961, pp. 95-107.

BIBLIOGRAPHY (Continued)

27. Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," Trans. ASME, Series D, Journal of Basic Engineering, 1960, pp. 35-40.

28. Kashyap, R. L., "Algorithms for Pattern Classification," in Adaptive Learning and Pattern Recognition Systems, J. M. Mendel and K. S. Fu, eds., Academic Press, New York, 1970.

29. Kishi, F. H., "On-Line Computer Control Techniques and Their Application to Re-entry Aerospace Vehicle Control," Advances in Control Systems, Vol. 1, C. T. Leondes, ed., Academic Press, New York, 1960.

30. Koford, J. S. and G. F. Groner, "The Use of an Adaptive Threshold Element to Design a Linear Optimal Pattern Classifier," IEEE Trans. Info. Theory, Vol. IT-12, January 1966, pp. 42-50.

31. Lawson, C. L. and R. J. Hanson, Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

32. Luenberger, D. G., Optimization by Vector Space Methods, John Wiley, New York, 1969.

33. Luenberger, D. G., Introduction to Linear and Nonlinear Programming, Addison Wesley, Reading, Massachusetts, 1973.

34. MacQueen, J., "Some Methods for Classification and Analysis of Multivariate Observations," Proceedings, Fifth Berkeley Symposium on Prob. and Stat., 1967, pp. 281-297.

35. Meisel, W. S., Computer-Oriented Approaches to Pattern Recognition, Academic Press, New York, 1972.

36. Minamide, N. and K. Nakamura, "A Restricted Pseudo Inverse and Its Application to Constrained Minima," SIAM Journal on Applied Mathematics, Vol. 19, No. 1, July 1970.

37. Moore, E. H., "On the Reciprocal of the General Algebraic Matrix," Abstract in Bulletin of American Mathematical Society, 1920, pp. 394-395.

38. Moore, E. H. General Analysis, Part I: Memoirs of the American Philos. Society, I, 1935, pp. 147-209.

39. Nagy, G., "State of the Art in Pattern Recognition," Proceedings IEEE, Vol. 56, No. 5, May 1968, pp. 836-861.

BIBLIOGRAPHY (Continued)

40. Nashed, M. Z., "Generalized Inverses, Normal Solvability, and Iteration for Singular Operator Equations," in Nonlinear Functional Analysis and Applications, L. B. Rall, ed., Academic Press, New York, 1971, pp. 311-359.

41. Nilsson, N. J., Learning Machines, McGraw-Hill, New York, 1965.

42. Noble, B., Applied Linear Algebra, Prentice Hall, Englewood Cliffs, New Jersey, 1969.

43. Noble, B., "Computational Methods for Generalized Inverses of Matrices and Related Results," in Generalized Inverses and Applications, M. Z. Nashed, ed., Academic Press, New York, 1975.

44. Papoulis, A., Probability, Random Variables, and Stochastic Processes, McGraw-Hill, New York, 1965.

45. Patterson, J. D. and B. F. Womack, "An Adaptive Pattern Classification System," IEEE Trans. Sys. Sci. Cyb., SSC-2, August 1966, pp. 62-67.

46. Penrose, R., "A Generalized Inverse for Matrices," Proceedings Cambridge Philos. Soc., Vol. 51, 1955, pp. 406-413.

47. Penrose, R., "On Best Approximate Solution of Linear Matrix Equations," Proceedings Cambridge Philos. Soc., Vol. 52, 1956, pp. 17-19.

48. Peters, G. and J. H. Wilkinson, "The Least-Squares Problem and Pseudo-Inverses," Computer Journal, Vol. 13, 1970, pp. 309-316.

49. Peterson, D. W. and R. L. Mattson, "A Method of Finding Linear Discriminant Functions for a Class of Performance Criteria," IEEE Trans. Info. Theory., Vol. IT-12, July 1966, pp. 380-387.

50. Pratt, William K., "Generalized Wiener Filtering Computation Techniques," IEEE Trans. on Computers, July 1972, pp. 636-641.

51. Pringle, R. M. and A. A. Rayner, Generalized Inverse Matrices with Applications to Statistics, Griffin, London, U. K., 1971

52. Rao, C. R. and S. K. Mitra, Generalized Inverse of Matrices and Its Applications, John Wiley, New York, 1971.

53. Sage, A. P., Optimum Systems Control, Prentice-Hall, Englewood Cliffs, New Jersey, 1968.

BIBLIOGRAPHY (Continued)

54. Sebestyen, G. D. and J. Edie, "An Algorithm for Nonparametric Pattern Recognition," IEEE Trans. Electronic Computers, Vol. EC-15, December 1966, pp. 908-915.

55. Sebestyen, G. D., Decision Making Processes in Pattern Recognition Macmillan, New York, 1962.

56. Shinozaki, N., et al., "Numerical Algorithms for the Moore-Penrose Inverse of a Matrix: Direct Methods," Annals of the Inst. of Stat. Math., Vol. 24, No. 1, 1972, pp. 193-203.

57. Shinozaki, N., et al., "Numerical Algorithms for the Moore-Penrose Inverse of a Matrix: Iterative Methods," Annals of the Inst. of Stat. Math., Vol. 24, No. 3, 1972.

58. Smith, F. W., "Design of Minimum-Error Optimal Classifiers for Patterns From Distributions with Gaussian Tails," IEEE Trans. Info., Theory, Vol. IT-17, No. 6, November 1971, pp. 701-707.

59. Smith, F. W., "Small-Sample Optimality of Design Techniques for Linear Classifiers of Gaussian Patterns," IEEE Trans. Info. Theory, Vol. IT-18, No. 1, January 1972, pp. 118-126.

60. Smith, S. E. and S. S. Yau, "Linear Sequential Pattern Classification," IEEE Trans. Info. Theory, Vol. IT-18, No. 5, September 1972, pp. 673-678.

61. Stewart, G. W., Introduction to Matrix Computations, Academic Press, New York, 1973.

62. Swerling, P., "Modern State Estimation Methods from the View Point of the Method of Least Squares," IEEE Trans. Aut. Control, Ac-16, December 1971, pp. 707-719.

63. Theil, H., Economic Forecasts and Policy, Second Edition, North-Holland Publishing Company, Amsterdam, 1961.

64. Theil, H., Principles of Econometrics, Wiley, New York, 1971.

65. Tou, J. T. and K. P. Heydron, "Some Approaches to Optimum Feature Extraction," in Computers and Information Sciences-II, J. Tou, ed., Academic Press, New York, 1967.

66. Van Trees, H. L., Detection, Estimation and Modulation Theory, Vol. 1, Wiley, New York, 1968.

BIBLIOGRAPHY (Concluded)

67. Watanabe, S., "Karhunen-Loéve Expansion and Factor Analysis," Theoretical Remarks and Applications," Information Theory, Statistical Decision Functions, Random Processes, Trans. 4th Prague Conf., 1965, pp. 635-640.

68. Wee, W. G., "Generalized Inverse Approach to Adaptive Multiclass Patterns Classification," IEEE Trans. Comput., Vol. C-17, December 1968, pp. 1157-1164.

69. Wee, W. G., "Generalized Inverse Approach to Clustering, Feature Selection, and Classification," IEEE Trans. Info. Theory, Vol. IT-17, No. 3, May 1971, pp. 262-269.

70. Wee, W. G., "On Feature Selection in a Class of Distribution-Free Pattern Classifiers," IEEE Trans. on Info. Theory, Vol. IT-16, No. 1, January 1970, pp. 47-55.

71. Wee, W. G. and K. S. Fu, "An Extension of the Generalized Inverse Algorithm to Multiclass Pattern Classification," IEEE Trans. on Systems Science and Cybernetics, July 1968, pp. 192-194.

72. Wilkinson, J. H. and C. Reinsch, Handbook for Automatic Computation, Vol. II, Linear Algebra, Springer-Verlag, New York, 1971.

73. Yau, S. S. and J. M. Garnett, "Least-Mean-Square Approach to Pattern Classification," in Frontiers of Pattern Recognition, M. S. Watanabe, ed., Academic Press, New York, 1972, pp. 575-587.

74. Young, T. Y. and T. W. Calvert, Classification, Estimation, and Pattern Recognition, American Elsevier, New York, 1974.

VITA

Mohamad Adnan Al-Alaoui was born in Damascus, The Syrian Arab Republic. He received a B.S. from Eastern Michigan University in August 1963, a B.S.E.E. from Wayne State University in December 1965, and a M.S.E.E. from the Georgia Institute of Technology in December 1968. He was an instructor at Detroit Engineering Institute from September 1964 to September 1965. He was an instructor at Radio, Electronics and Television Schools in Detroit, Michigan, from June 1965 to February 1966. He was an Assistant Project Engineer at Bendix Radio Division in Baltimore, Maryland, where he worked in the Avionics Department from February 1966 to September 1967. He served as a teaching assistant in the Electrical Engineering Department of the Georgia Institute of Technology from September 1967 to September 1974. He also served concurrently as a teaching assistant in the School of Mathematics at the Georgia Institute of Technology from September 1971 to June 1972.